Inferring biological tasks using Pareto analysis of high-dimensional data

Yuval Hart[1,3], Hila Sheftel[1,3], Jean Hausser[1,3], Pablo Szekely[1,3], Noa Bossel Ben-Moshe[2], Yael Korem[1], Avichai Tendler[1], Avraham E Mayo[1] & Uri Alon[1]

We present the Pareto task inference method (ParTI; http://www.weizmann.ac.il/mcb/UriAlon/download/ParTI) for inferring biological tasks from high-dimensional biological data. Data are described as a polytope, and features maximally enriched closest to the vertices (or archetypes) allow identification of the tasks the vertices represent. We demonstrate that human breast tumors and mouse tissues are well described by tetrahedrons in gene expression space, with specific tumor types and biological functions enriched at each of the vertices, suggesting four key tasks.

Approaches for analyzing high-dimensional data sets[1–5] include dimensionality reduction techniques such as principal-component analysis (PCA)[6], *t*-distributed stochastic neighbor embedding (t-SNE)[7] and methods that split data points into groups, such as clustering[8] and Gaussian mixture models (GMMs)[8]. A recent advance suggests a complementary way to understand large biological data sets on the basis of Pareto optimality of biological systems with respect to multiple evolutionary tasks[9–11]. Here we present a method for identifying such tasks.

The Pareto approach notes that cells that need to perform multiple tasks face a fundamental trade-off: no gene expression profile can be optimal for all tasks faced by the cell. Shoval *et al.*[9] showed that the best compromises between tasks lead to phenotypes that lie in low-dimensional polytopes in trait space (for example, gene expression space). Two tasks lead to points arranged along a line, three tasks to a triangle, four tasks to a tetrahedron and so on. The vertices of these polytopes are called archetypes, and they represent the optimal phenotype for a single task[9] (**Supplementary Note 1** and **Supplementary Fig. 1**).

If a biological data set represents a Pareto-optimal situation, then (i) it should fall inside a polytope, and (ii) the points nearest each vertex of the polytope (each archetype) should correspond to specific biological tasks or functions. In other words, key biological features related to a given task should be maximally enriched near the archetype that corresponds to that task.

To implement these ideas, we present the ParTI method (http://www.weizmann.ac.il/mcb/UriAlon/download/ParTI). The input is a set of N data points, described as vectors of K traits. Each data point is also annotated by a vector of M additional features. The method has two stages: (i) compute the minimal polytope that encloses the data and its statistical significance and (ii) compute the enrichment of each feature as a function of the distance from each archetype. The data are well explained by ParTI if a low-dimensional polytope significantly describes them, and if there are features that are maximally enriched near each archetype (**Supplementary Note 2** and **Supplementary Fig. 2**).

We determine the number of archetypes by fitting polytopes with n vertices to the data using principal convex hull analysis (PCHA)[12]. We choose n beyond which there is little improvement in the explained variance (EV), according to an 'elbow' test (Online Methods). Archetype positions are then determined by hyperspectral unmixing algorithms[13–16]. Statistical significance is assessed by the *t*-ratio test[9] (Online Methods). These algorithms have rarely been applied to biological data, with the notable exceptions of studies by Schwartz and Shackney[17] and Tolliver *et al.*[18] that analyzed tissue mixtures in tumors and research by Thøgersen *et al.*[19] that used PCHA to analyze bacterial gene expression.

The second stage of the method provides clues for the biological task that is optimized at each archetype (**Fig. 1a**). A simple approach is to characterize each archetype by searching for gene sets over-represented among the genes differentially expressed at the archetype[19]. We find that this can be improved by considering the density of each feature as a function of distance from the archetype: we seek features (for example, Molecular Signature Database (MSigDB)[20] gene-set expression levels and clinical information) whose density peaks at the points closest to the archetype. For this purpose, we bin data using a computed optimal bin size and determine which feature is maximal in the bin closest to an archetype. Statistical significance of this maximization and relevant multiple-hypotheses testing controls (false discovery rate (FDR) and randomization tests) are computed (Online Methods). We use a leave-one-out strategy to avoid circularity concerns when a feature (for example, gene) is also used to define the archetype location.

We demonstrated ParTI on a gene expression data set of 2,106 human breast cancer tumors and healthy tissues[3]. These data reside in a 6,970-dimensional gene expression space and are annotated by a vector of 181 clinical features (**Supplementary Table 1**). A tetrahedron (four-vertex polytope) best described the data ($P < 10^{-4}$; **Fig. 1b**). Projections on facets resembled triangles (**Fig. 1b**). Vertex coordinates lay well outside of the data (**Fig. 1b**). The **Supplementary Results** and **Supplementary Figure 3** provide more details.

On average, 12 features (of 181) were maximally enriched at each archetype (FDR < 0.1). The enriched features corresponded

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel. [3]These authors contributed equally to this work. Correspondence should be addressed to U.A. (urialon@weizmann.ac.il).

RECEIVED 11 JULY 2014; ACCEPTED 17 NOVEMBER 2014; PUBLISHED ONLINE 26 JANUARY 2015; DOI:10.1038/NMETH.3254

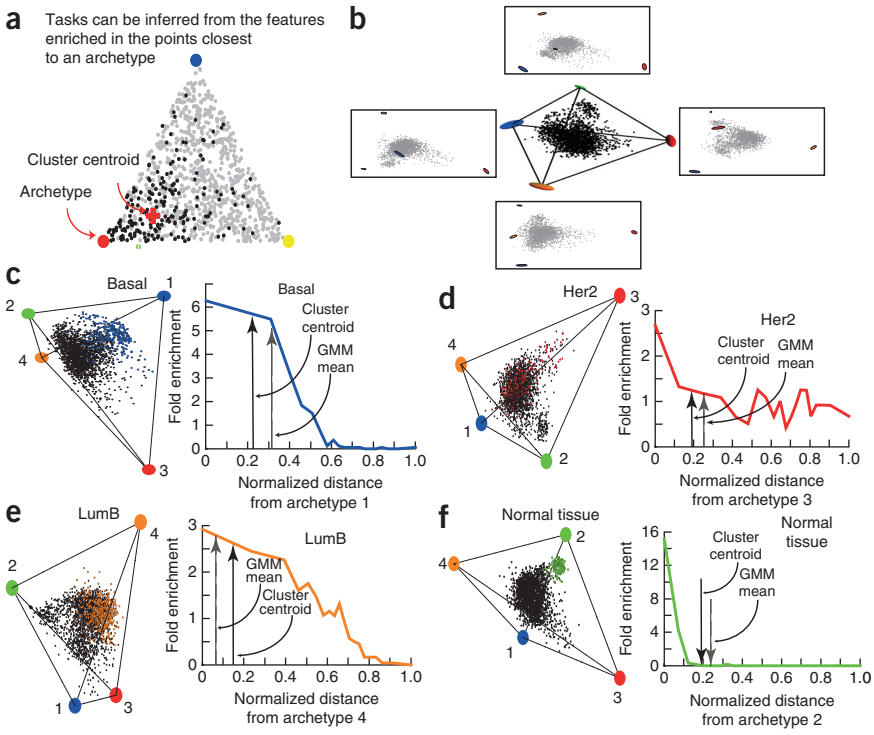

Figure 1 | Key cancer features are maximally enriched at points nearest the archetypes. (**a**) ParTI tests for the maximal enrichment of features near the archetypes. Note that maximal enrichment in this case is not at the cluster centroid (plus sign). (**b**) Three-dimensional (3D) plot of the data and enclosing tetrahedron. The axes are the first three principal components. The colored ellipsoids represent the archetype location and error on the most varying directions. Archetype error bars are obtained by bootstrapping. Each ellipsoid represents 68% confidence level. The inset near each archetype shows the projection of the data on the plane defined by the tetrahedron's face opposing that archetype. (**c**–**f**) 3D plots of basal tumor (**c**), Her2 (**d**), LumB (**e**) or normal tissue (**f**) in the breast cancer data set. Numbers indicate the archetypes. 2D plots show enrichment as function of rank order of bins of 5% of the points sorted by Euclidean distance from archetype. The positions of the centroid of the cluster and the Gaussian mixture model (GMM) mean corresponding to each feature are shown by arrows. (**g**) Inferred tasks of each archetype in the breast cancer data set, along with PAM50 subtype (Online Methods), representative clinical features and gene sets enriched in the vicinity of each archetype. For complete lists, see **Supplementary Table 2**.

g

Archetype	Inferred tasks	PAM50 subtype	Enriched clinical features and gene sets
1	De-differentiation and division	Basal	Poorly differentiated cells, ESR1$^-$, PGR$^-$, ERBB2$^-$, proliferation
2	Differentiation in healthy tissue	Normal	Normal tissue, lipid metabolism, adipocyte differentiation, negative regulation of cell cycle
3	Signaling	Her2	Protein kinase activity, Notch, Erbb2, JNK and GnRH signaling, protein import into nucleus
4	Growth	LumB	Translation, phospholipid biosynthesis, ATP synthesis

to major clinical subtypes. Archetype 1 was maximally enriched with samples classified as basal tumors, which are high-grade ESR1$^-$ERBB2$^-$PGR$^-$ tumors. Archetype 2 corresponded to normal tissue samples. Archetype 3 was enriched for tumors of the Her2 subtype. Archetype 4 corresponded to LumB tumors, which are mostly ESR1$^+$ERBB2$^-$ tumors in the present data set. These features showed maximal enrichment at the points closest to the archetype rather than at the centroid of the clusters obtained by *k* means or by GMM (**Fig. 1c–f**, **Supplementary Tables 2–5**, **Supplementary Notes 3–5** and **Supplementary Fig. 4**).

We also found biological functions (MSigDB[20] gene sets) maximally enriched at the points closest to each archetype (Online Methods, **Supplementary Tables 2** and **6**). These gene sets fit well with the biology of cancer subtypes enriched at the archetypes (**Supplementary Discussion**). Comparing the gene-set enrichment using ParTI to the method of ref. 19 (**Supplementary Table 7**), we found statistical significance (*P* values) typically improved by about 20 orders of magnitude and improved gene-set coherence (one edge closer on a Gene Ontology tree on average), yielding a set of more related and hence more interpretable biological functions (**Supplementary Note 6** and **Supplementary Fig. 5**).

Enriched clinical features and gene sets suggested putative tasks for three of the archetypes: for archetype 1, cell de-differentiation and division; for archetype 2, cell differentiation; and for archetype 4, high metabolism and growth. Archetype 3 was harder to understand: it was enriched with signaling pathway expression (**Fig. 1c–g**). Pareto theory suggests that the trade-off between these tasks shapes the distribution of tumors in expression space. For example, differentiation and de-differentiation are tasks that cannot be achieved at the same time. An alternative explanation is that ParTI detects the relative proportion of different cell types in each tissue sample. However, we found normal tissues enriched at only one archetype despite their different breast contexts, hinting that composition may not be dominant in the present context. The inferred tasks may provide clues for effective therapy targets. For example, the tumors near the rapid-growth archetype (4) could be affected by therapy that blocks metabolic growth pathways. The division archetype (1) suggests drugs that target dividing cells.

We extended the enrichment analysis also to higher numbers of archetypes. Taking more than four archetypes resulted in a split of the basal tumor archetype into several archetypes, leaving the other archetypes nearly unchanged (**Supplementary Results** and **Supplementary Fig. 6**). One may consider these as potential subtypes of basal breast cancer, a tumor category thought to have at least six clinical subtypes[21].

An mRNA-Seq data set of 1,106 breast tumors[5] yielded similar results (**Supplementary Table 8**): data were well described by a tetrahedron, and three of the four archetypes (1,2 and 4) matched the archetypes found above in terms of enriched gene sets and clinical types (**Supplementary Results** and **Supplementary Fig. 7**). This further supports the inference of key tasks for three of the archetypes.

We also analyzed 4,364 genes expressed in at least 95% of 63 mouse tissues[22] (**Supplementary Note 7**). These genes are expressed in most cell types and thus do not include the marker genes that are often used to define cell types. Most of the 63 tissues are thought to each be made of a single cell type. We find that the data are best explained by four archetypes ($P < 0.01$, **Fig. 2**). The archetypes were close to bone marrow macrophage

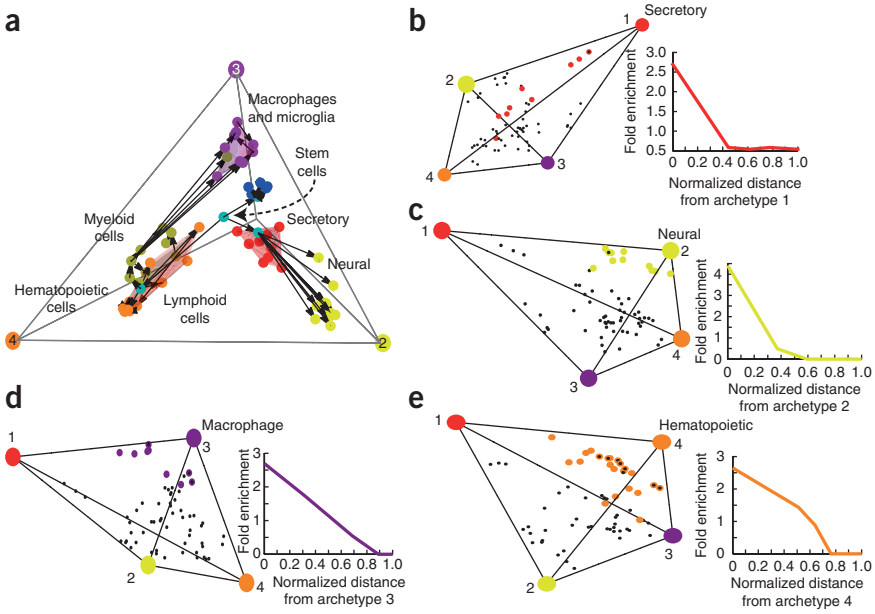

Figure 2 | A mouse tissue gene expression data set is well described by a tetrahedron, with archetypes enriched with specific features. (**a**) Embryonic stem cells (light blue) are at the center of the tetrahedron. As they differentiate, they come closer to the facets. Arrows represent differentiation into neural (yellow), hematopoietic (orange for lymphoid, olive green for myeloid) and macrophage cells (purple). (**b**–**e**) Three-dimensional plots of enrichment near each archetype and two-dimensional plots of enrichment as function of rank order of bins of 20% of the points sorted by Euclidean distance from each archetype.

(archetype 3, macrophage and microglia cells), $CD4^+CD8^+$ T cells (archetype 4, lymphoid cells), amygdala (archetype 2, neural cells) and pancreas (archetype 1, secretory glands). Enrichment analysis suggested specific functions for the archetypes: locomotion, digestion and cell-cell communication (3); proliferation and antigen presentation (4); communication across synapses (2); and secretion (1) (**Fig. 2** and **Supplementary Table 9**). Indeed, these functions are hallmarks of the respective archetypal tissues. This suggests tasks that may trade off in the rest of the tissue types: for example, locomotion and enzymatic digestion in the macrophages and microglia might not be feasible together with rapid cell division as in lymphoid and progenitor cells.

The tetrahedron was nonuniformly populated; differentiated tissues mostly occupied one triangular facet between archetypes 1, 3 and 4. The remaining differentiated tissues, namely all neural tissues, were along the edge from archetype 2 to archetype 1 (**Supplementary Fig. 8**). Thus, two potential trade-offs—the two edges between archetypes 2 and 3 and archetypes 2 and 4—were not found. Embryonic stem cells were near the center of the tetrahedron, as expected for generalists. Stem cells on their way to differentiation (hematopoietic stem cells) came closer to the face of the tetrahedron as they approached their differentiated fate (**Fig. 2a**).

The analyzed data sets are cell-population averages for each data point; but one may apply ParTI also to analyze single-cell data to study the variation among individual cells in a population. ParTI has caveats: a data set may resemble a polytope owing to reasons unrelated to Pareto optimality—for example, if experimental error increases with trait magnitude or when outliers dictate an archetype (**Supplementary Fig. 9**). Many of these cases can be resolved by archetype enrichment analysis (**Supplementary Discussion**). Further work can improve the method's statistical power and efficiency. Analyzing more biological data sets can help to determine the biological prevalence of polytopes induced by trade-offs between tasks.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank N. Drayman, B. Towbin, M. Botzman, Y. Liron, M. Adler, G. Aidelberg, D. Rothschild, S. Malihi, O. Szekely and members of the Alon lab for discussions.

We acknowledge support by the Human Frontier Science Program, project number RGP0020/2012, European Research Council, project number 249919, and Rising Tide Cancer Research Fund, project number 721176. U.A. receives support as the Abisch-Frenkel Professorial Chair. J.H. acknowledges the support of the Swiss National Science Foundation (PBBSP3_14961) and EMBO (ALTF 1160-2012).

AUTHOR CONTRIBUTIONS

Y.H., H.S., J.H. and P.S. developed the method and analyzed the data. N.B.B.-M. analyzed the microarray breast cancer data. Y.K., A.T. and A.E.M. consulted on the method and algorithm. U.A. designed the method and research program. Y.H., H.S., J.H. and P.S. wrote the Matlab code, and Y.H., H.S., J.H., P.S. and U.A. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Kim, H.D., Shay, T., O'Shea, E.K. & Regev, A. *Science* **325**, 429–432 (2009).
2. Kalisky, T., Blainey, P. & Quake, S.R. *Annu. Rev. Genet.* **45**, 431–445 (2011).
3. Curtis, C. *et al. Nature* **486**, 346–352 (2012).
4. Bendall, S.C. & Nolan, G.P. *Nat. Biotechnol.* **30**, 639–647 (2012).
5. The Cancer Genome Atlas Network. *Nature* **490**, 61–70 (2012).
6. Ringnér, M. *Nat. Biotechnol.* **26**, 303–304 (2008).
7. Van der Maaten, L. & Hinton, G. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
8. Hastie, T., Tibshirani, R. & Friedman, J. in *The Elements of Statistical Learning* 2nd edn. 520–528 (Springer, 2009).
9. Shoval, O. *et al. Science* **336**, 1157–1160 (2012).
10. Sheftel, H., Shoval, O., Mayo, A. & Alon, U. *Ecol. Evol.* **3**, 1471–1483 (2013).
11. Szekely, P., Sheftel, H., Mayo, A. & Alon, U. *PLoS Comput. Biol.* **9**, e1003163 (2013).
12. Mørup, M. & Hansen, L.K. *Neurocomputing* **80**, 54–63 (2012).
13. Li, J. & Bioucas-Dias, J.M. *IEEE Int. Geosci. Remote Sens. Symp.* **3**, 250–253 (2008).
14. Chan, T.-H., Chi, C.-Y., Huang, Y.-M. & Ma, W.-K. *IEEE Trans. Signal Process.* **57**, 4418–4432 (2009).
15. Chan, T.-H., Liou, J.-Y., Ambikapathi, A., Ma, W.-K. & Chi, C.-Y. in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 1237–1240 (IEEE, 2012).
16. Bioucas-Dias, J.M. *et al. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **5**, 354–379 (2012).
17. Schwartz, R. & Shackney, S.E. *BMC Bioinformatics* **11**, 42 (2010).
18. Tolliver, D., Tsourakakis, C., Subramanian, A., Shackney, S. & Schwartz, R. *Bioinformatics* **26**, i106–i114 (2010).
19. Thøgersen, J.C., Mørup, M., Damkiær, S., Molin, S. & Jelsbak, L. *BMC Bioinformatics* **14**, 279 (2013).
20. Subramanian, A. *et al. Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
21. Lehmann, B.D. *et al. J. Clin. Invest.* **121**, 2750–2767 (2011).
22. Lattin, J.E. *et al. Immunome Res.* **4**, 5 (2008).

ONLINE METHODS

Calculating the explained variance and determining the number of archetypes. Dimensionality of the data is first reduced by principal-component analysis (PCA) down to eight dimensions (using 20 dimensions provides very similar results). For each number of archetypes n, we find the best-fit polytope using the PCHA algorithm[12] with $\delta = 0$. We compute the explained variance given by the mean relative distance of the N data points to the polytope

$$\mathrm{EV}(n) = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \|\vec{p}_i - \vec{s}_i\| / \|\vec{p}_i\|\right)$$

Here p_i is the ith data point and s_i is the closest point to p_i in the polytope[19,23]. For points inside the polytope,

$$\|\vec{p}_i - \vec{s}_i\| = 0$$

We seek a number of archetypes for which adding an additional archetype does not increase EV by much. Operationally, we seek the value of n at which the EV(n) curve has a bend; typically, the EV curve has a rapid rise with n that switches to a slower rise at higher values of n. The value of n at which the bend occurs is estimated by finding the 'elbow' of the curve (**Supplementary Fig. 10**).

This method for determining the data dimensionality differs from PCA. PCA can find the dimensionality of the data set and provides a set of orthogonal axes along which the data vary most. PCA, however, does not indicate whether the data lie in a polytope. We used PCA as a pre-step to reduce data dimensionality. Using PCA is not essential, as some polytope-finding algorithms also work well in the full original high dimensionality of the data set. More subtly, the orientation of the archetypes does not generally align with the first principal components. This implies that the biological meaning of the archetypes is distinct from that carried by the PCA components.

Estimating archetype position. After determining the number of archetypes, we use a hyperspectral unmixing algorithm to find the archetype positions. The software includes settings that allow use of one of five different algorithms. To fit the data to a polytope, we start by performing PCA on the data (thus centering the data to have a zero mean), without normalizing the data by their s.d. (that is, without Z-scoring). For the case of n archetypes, we use data projected on the first $n - 1$ principal components (i.e., for a tetrahedron, we use the first three components to represent the data). Then we use an unmixing algorithm to fit the n-vertices polytope that best describes the data (see **Supplementary Note 8** and **Supplementary Fig. 11** for a list of algorithms). We used Sisal[24] for the cancer data and MVSA[13] for tissues because MVSA does not allow outliers and thus is more appropriate for a small number of data points.

Archetypal analysis differs from clustering analysis and Gaussian mixture models (GMMs) in several aspects—namely, a graded score for each point according to its distance from the archetypes rather than a discrete assignment into clusters, and lower sensitivity to local dense clumps that may result from data sampling. For a general comparison of these approaches, see Mørup and Hansen[12] and **Supplementary Note 2**.

When using unmixing algorithms such as Sisal or MVSA, the archetypes are usually located some distance outside of the data. In this case, they represent hypothetical gene expression profiles that, according to the theory, should correspond to optimal profiles for the different tasks (or to archetypal cell mixtures in the case of tissue heterogeneity). We computed error bars on the archetypes by resampling the data with replacement and computing the archetypes 1,000 times (see **Supplementary Note 9** and **Supplementary Fig. 12**). The error in estimating the archetypes is about 10% (s.d./mean). The s.d. values of the archetype positions are depicted as ellipsoids in **Supplementary Figure 13** for the mouse tissue data set and in **Figure 1b** for the breast cancer data set.

Evaluating significance of best-fit polytopes. We note that estimating the number of archetypes using EV curves suggests the number of vertices of the best-fit polytope but does not mean that the data are necessarily well fit by a polytope (**Supplementary Discussion** and **Supplementary Fig. 14**).

To estimate the statistical significance (P value) of the description of the data by an n-vertex polytope, we compute a measure for the extent that the data fill the polytope, known as the t-ratio. The t-ratio is defined as the ratio of the volume of the polytope to the volume of the convex hull of the data[9]. This ratio is usually larger than 1; the closer it is to 1, the better the enclosing polytope captures the shape of the data. We then generate randomized data sets where for each point, values of each trait are sampled independently from its ensemble of measured values. This preserves the distribution of values of each trait while eliminating correlations between traits. We calculate the t-ratios for each randomized data set and set the P value to be the proportion of randomized sets with a t-ratio smaller than or equal to that of the original data.

The statistical significance of the fitted polytope depends both on the dimension of the data set and the number of data points. Generally we find that a few tens of data points are sufficient for preliminary analysis (data not shown).

Evaluating the enrichment of features for each archetype. We seek those features that are maximally enriched at the points nearest each archetype. Each data point is associated with a value for each of M features. A feature can be categorical, for example, the PAM50 feature (a computational classification of tumor subtypes based on gene expression profiling) in the breast cancer data set that can only take specific values (basal, LumB, …). We transform such categorical features into Boolean features (true or false). For example, the PAM50 feature becomes a set of Boolean features (PAM50-basal, PAM50-LumB, …). Other features are continuous, for example, patient age. We begin by defining the density profile for each feature as a function of distance from the archetype. For this purpose, we sort points in increasing order of Euclidean distance from archetype i. We bin all points in the data set according to their distance from the given archetype, such that each bin has an equal number of points (see "Calculating the optimal bin size" below). We compute the enrichment of feature j in the bins of sorted points. For discrete features, enrichment is defined as the density of the feature in the bin relative to its mean density across all data. To calculate the significance of the enrichment in the bin closest to the archetype, we use the hypergeometric test.

doi:10.1038/nmeth.3254

A continuous feature can be treated in two ways. First, one can bin the feature values into several bins and then treat them as a categorical variable (this was used for features with low information content; see **Supplementary Note 10**). A second option is to define enrichment as the median value of the property in the bin and to use continuous significance tests such as the Mann-Whitney test[25] for calculating the P value. This method finds enrichment for only high or low feature values and not for intermediate values.

We plot enrichment curves as a function of the median distance of each bin to the archetype and normalize the binned distance between 0 and 1.

In cases where many features are tested for enrichment, one must control for multiple-hypothesis testing. We used two methods: one is a standard false discovery rate (FDR) calculation[26] (with a threshold of 0.1, in our case). The second controls for multiple-hypothesis testing by comparison to data sets in which the feature vectors are randomly shuffled between data points (**Supplementary Note 11** and **Supplementary Fig. 15**). Both tests provide similar results and suggest that multiple-hypotheses errors are minimal for both data sets.

We record the features whose maximal enrichment is at the bin closest to the archetype. To do so, we calculated the probability that the fraction of data points with a certain feature is maximally enriched in the bin closest to the archetype. We first determined, in each bin, the probability distribution of the fraction of points with the feature, which follows a beta distribution $B(m+1, q+1)$, where m and q are the number of points in the bin with and without the feature, respectively. We then computed the probability that this fraction is higher in the first bin than in all other bins. Taking the product of the probabilities that the fraction of points with the feature is larger in bin 1 than in all other bins yields the probability $P_{\max}$ that the feature is maximally enriched close to the archetype (see **Supplementary Note 12** for the detailed derivation). This is a stringent criterion because it removes features that peak some distance away from the archetype (**Supplementary Note 6**). The latter features cannot be associated with the tasks and trade-offs at play according to Pareto theory[9,10]. Moreover, we use a leave-one-out strategy to avoid circularity concerns that stem from using a trait for both defining the archetype and measuring its enrichment. Finally, in our approach, no threshold on gene regulation needs to be chosen beyond which to look for over-represented genes sets.

Calculating the optimal bin size. We developed an approach to determine the bin size for enrichment analysis. We seek the minimal bin size in which stochastic effects do not mask out the signal. To estimate the minimal bin size, we consider the following simple calculation. The total number of data points is N, and m of them have a given feature, so that the mean density is $\rho_0 = m/N$. We assume that the feature density $\rho(x)$ is maximal at an archetype and decreases with distance from the archetype, such that to a first-order approximation

$$\rho(x)/\rho_0 = (1 - x/\lambda)\varphi$$

Here φ is the enrichment at the archetype, and λ is the length scale over which enrichment decays from the archetype (related to the first derivative of $\rho(x)$). Consider bins of size b (as a fraction of total number of points). The number of feature points in the first bin (closest to archetype) is $n_1 = \varphi\rho_0 Nb$. The number of feature points in the second bin is $n_2 = \varphi(1 - b/\lambda)\rho_0 Nb$. Thus, the difference between the first and second bin is

$$\Delta = \varphi b^2\ \rho_0 N/\lambda$$

We seek a bin size such that the s.d. due to random fluctuations in the two bins is smaller than the expected difference Δ. The number of points with a given feature value at each bin is given by a binomial process, with probability p for a feature point. For the first bin: $p = \rho(0) \approx \varphi\rho_0$, and the variance is given by $\sigma_1^2 \approx \sigma_2^2 \approx p(1-p)Nb$. We require $\sigma_1 + \sigma_2 < \Delta$. This results in the inequality

$$b > \left(4\frac{(1-\varphi\,\rho_0)\lambda^2}{\varphi\rho_0 N}\right)^{\frac{1}{3}}$$

The power laws result from the fact that bin size b has two effects: it controls the number of points in each bin and also the difference between the first and second bin. The 1/3 power law means that the minimal bin size depends only weakly on N, ρ_0 and φ and depends most strongly on λ, the rate of decay of the enrichment with distance from the archetype. For the cancer data set, we find that the key features require an estimated minimal bin size of 1–5% (the four features in **Fig. 1c–f** show $\lambda = 0.25, 0.27, 0.45, 0.17$; $\varphi = 6.3, 2.7, 2.9, 15.1$; $\rho_0 = 0.16, 0.11, 0.23, 0.06$; and $N = 2{,}106$, respectively, resulting in $b_{\min} \approx 1\%, 7\%, 6\%, 2\%$). For simplicity, we use 5% bins in all calculations in the main text.

To further test the minimal bin size needed, we reanalyzed the cancer data set with bin sizes ranging from 1% to 30% in steps of 1%. At each bin size, we used bootstrapping to generate 100 resampled data sets (generating N data points by sampling with replacement). We asked at which bin size the P value for enrichment varies the least while resulting in the most significant P values. We considered four enriched features: basal, normal tissue, Her2 and LumB, for archetypes 1, 2, 3 and 4, respectively. This yields minimal bin size estimates of 5–10%.

Gene-set enrichment analysis. We test whether specific gene functions and pathways are significantly enriched close to the archetypes. To do so, we use MSigDB database[20] to define sets of genes belonging to the same pathway or sharing a common biological function. Specifically, we use the c2.cp.v4.0.symbols.gmt file, which contains curated gene sets from canonical pathways annotated by BioCarta[27], KEGG[28] and Reactome[29]. We also use the c5.all.v4.0.symbols.gmt file, which is based on the Gene Ontology (GO): there, genes annotated with the same GO term are grouped into the same gene set. We compute the expression level of each gene set by averaging over the $\log_2$ expression of all genes in that gene set. We discard gene sets that contain fewer than ten expressed genes in the gene expression data set. By repeating this procedure for all samples and all gene sets, we obtain a matrix in which cells represent the amount of expression of a gene set for each sample. We use a Mann-Whitney test to determine which gene sets are significantly overexpressed in the samples closest to each archetype compared to all other samples in the data set. Only gene sets with an FDR[26] smaller than 0.1 are kept for further

analysis. Pareto theory implies that the enrichment of properties which correspond to archetypical tasks should peak at the archetype. We can therefore eliminate false positives by considering only gene sets whose median expression is highest in the bin closest to the archetype. By repeating this procedure for all gene sets, we obtain a list of gene sets significantly enriched close to each archetype. To address circularity concerns stemming from using gene expression both to infer the position of the archetypes and their tasks, we use a leave-one-out procedure: for each enriched gene set, we recompute the position of the archetypes after removing the genes in that gene set. We then determine which samples are closest to the new archetypes, and test whether the gene set is still significantly enriched close to the archetype by the same method as above (Mann-Whitney test[25]). Finally, we rank the resulting list of significantly enriched gene sets by the difference between their median expression in the bin closest to each archetype and their median expression in all other samples.

Code availability. The entire analysis is implemented as a Matlab software package available from http://www.weizmann.ac.il/mcb/UriAlon/download/ParTI.

23. Cutler, A. & Breiman, L. *Technometrics* **36**, 338–347 (1994).
24. Bioucas-Dias, J.M. in *Hyperspectral Image Signal Process. Evol. Remote Sens. First Workshop* 1–4 (IEEE, 2009).
25. Mann, H.B. & Whitney, D.R. *Ann. Math. Stat.* **18**, 50–60 (1947).
26. Benjamini, Y. & Hochberg, Y. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
27. Nishimura, D. *Biotech Softw. Internet Rep.* **2**, 117–120 (2001).
28. Kanehisa, M. & Goto, S. *Nucleic Acids Res.* **28**, 27–30 (2000).
29. Croft, D. *et al. Nucleic Acids Res.* **39** (suppl. 1), D691–D697 (2011).

doi:10.1038/nmeth.3254