Deciphering causal genomic templates of complex molecular phenotypes

Salil S. Bhate[1]*, Anna Seigal[2], and Juan Caicedo[1]

1. Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA, 02142 USA.
2. Harvard University, Cambridge, MA, 02138 USA.

*Correspondence: sbhate@broadinstitute.org

Abstract

We develop a mathematical theory proposing that complex molecular phenotypes (CMPs, e.g., single-cell gene expression distributions and tissue organization) are produced from templates in the genome. We validate our theory using a procedure termed Causal Phenotype Sequence Alignment (CPSA). CPSA finds a candidate template of a CMP by aligning – without using genetic variation or biological annotations – a phenotypic measurement (e.g., a tissue image) with a reference genome. Given any edit to the CMP (e.g., changing cellular localization), CPSA outputs the genomic loci in the alignment corresponding to the edit. We confirm that three CMPs (single-cell gene expression distributions of the immune system and of embryogenesis, and tissue organization of the tumor microenvironment) have templates: the loci output by CPSA for therapeutically significant edits of these CMPs reveal genes, regulatory regions and active-sites whose experimental manipulation causes the edits. Our theory provides a systematic framework for genetically redesigning CMPs.

Introduction

Technological advances have enabled measurement of complex molecular phenotypes (CMPs) like single-cell gene expression distributions (measured by sequencing (Kolodziejczyk et al., 2015)) and tissue organization (measured by high-parameter imaging (Palla et al., 2022)). A biological challenge is therefore to understand how CMPs are causally specified by the genome (Palla et al., 2022). A causal understanding would enable redesigning CMPs using genetic perturbations (Cong et al., 2013; Qi et al., 2013) as well as discovering genetic targets for achieving therapeutically important edits to CMPs.

The current understanding of how the genome specifies CMPs is via emergence: regulated expression of and interactions between proteins and mRNAs encoded in the genome give rise to the CMPs observed in high-parameter datasets (**Figure 1A.1 the blue structure is the CMP produced by the genome, represented as black box**). Genome-wide association-studies (Schmiedel et al., 2022; Walker et al., 2022; Yazar et al., 2022), functional-genomics screens (Przybyla and Gilbert, 2022) and ex-vivo reconstitution approaches (Hodis et al., 2022) reveal the specific genomic loci associated with specific CMPs. However, such approaches do not provide a theoretical understanding of how the genome specifies CMPs in generality. In addition, these genome-wide approaches can only be applied when CMPs can be measured at scale in patient cohorts (unavailable in the case of rare diseases (Wangler et al., 2017)) or recapitulated ex vivo (not always possible for tissue organization and multicellular interactions (Rossi et al., 2018)).

In contrast to CMPs, an established theory explains how amino-acid sequences of proteins are causally specified by the genome: any amino-acid sequence has a corresponding coding-sequence in the organism's genome (Crick, 1970). The coding-sequence can be regarded as a "template": achieving expression of an edited protein requires only finding the template in the genome and making the corresponding edit to the template. This theory has practical applications: any prescribed edit to a protein can be introduced from only knowledge of its

sequence and the reference genome, without measurements of genetic or phenotypic variation or ex vivo recapitulation of the protein's translation.

The formal theory that protein sequences have templates in the genome required three steps. First, the notion of a template of a protein within a genome as a coding-sequence had to be mathematically formulated. Second, a search algorithm (e.g., the BLAST algorithm (Altschul et al., 1990)) was required to find candidate templates of protein sequences in genome sequences – in this case, a candidate template is a local alignment of a protein sequence within a genome. Third, it had to be experimentally verified that, given any prescribed edit to a protein sequence, genetically manipulating the corresponding loci in its template resulted in expression of the edited protein.

Like protein sequences, CMPs are structures of extreme combinatorial complexity that are reliably reproduced by organisms (**Figure 1A.1 the blue structure is the CMP output by the genome**). We therefore hypothesized that CMPs are produced according to templates within the genome. The practical implication of this hypothesis is that, if correct, an edit to a CMP can be genetically designed from only a phenotypic measurement and the reference genome of the organism producing the CMP. How might it be assessed whether CMPs had such templates?

First, the notion of a template of CMP within a genome must be mathematically formulated. Second, a search algorithm is required to identify a candidate template of a given CMP within the reference genome of the organism producing the CMP **(Figure 1A.2, black structure of identical shape to CMP in genome**). Finally, it must be experimentally verified whether genetically manipulating (e.g., with genome-editing) the genomic loci in the candidate template corresponding to any prescribed edit to the CMP (**Figure 1A.3, prescribed edit of CMP is on top row, corresponding edit of template on bottom row**) results in expression of the correctly edited CMP (**Figure 1A.3, bottom row, red arrow outputs the correctly edited CMP**). A **Glossary** is provided in the **Methods** as an informal reference for the new terms introduced.

We first developed a mathematical notion of template that could be applied to CMPs. A template of a CMP would not be a contiguous sequence in the genome (as seen for protein sequences) for two reasons. Mathematically, CMPs are not sequences, but rather, complex, high-dimensional structures. Biologically, loci from across the genome are involved in regulation of CMPs. Our definition addresses these two concerns and involves reformulating the genome as a metric space (instead of as a sequence). The definition is described in **Figure 1** and **Supplementary Note 1, Section 1**. The definition is formally analogous to defining a template of a protein sequence via deciphering the amino-acid code, as we detail in **Supplementary Note 2**.

Secondly, we formulated the identification of candidate template of a CMP as a search problem: to align a measurement of a CMP (e.g. a multiplexed immunofluorescence image) with a reference genome sequence (e.g., GRCh38). The resulting alignment – referred to as a "phenotype-sequence alignment" (PSA) **–** is then the candidate template (**Figure 1A.2, black structure of identical shape to CMP in genome**). Like the coding-sequence of a protein in a genome, a PSA is a formal copy of the information in a CMP in a genome.

We designed and implemented a neural network optimal transport algorithm termed "Neural Alignment of CMP with Reference genome" (NACR) to find PSAs. The only two inputs to NACR are a reference genome sequence and a measurement of a CMP, and the output is a PSA. Genetic variation, external data or biological annotations are not utilized. The NACR algorithm and validation that it, without a priori incorporation of any genetic information beyond the reference genome, outputs PSAs capturing specific genetic information specific to the input CMPs is presented in **Figure 2** of the **Results**; the algorithm is formally defined in **Supplementary Note 1, Section 2.**

The third and final step is to experimentally verify that a PSA of a CMP is a template of the CMP. This requires: a) selecting edits to the CMP; b) finding the genomic loci in the PSA corresponding to the edits and c) confirming that experimental manipulation of these genomic

loci results in correctly edited CMP. We developed a procedure implementing (b) termed "causal phenotype sequence alignment" (CPSA). CPSA takes as input a CMP and a PSA with a reference genome, and maps in silico edits to the CMP (**Figure 1A.3, prescribed edit of CMP is on top row**) to the genomic loci in the PSA corresponding to the edits (**Figure 1A.3, corresponding edit to PSA on bottom row**). These loci are produced by CPSA without biological annotations, external data or measurements of genetic variation. If (c) can be can be successfully experimentally verified using the genomic loci output by CPSA, we consider the PSA to be a template.

Why might CMPs have templates as revealed by our approach? We show using simulations in **Figure 3** of the **Results** that under information-efficiency constraints on the genomic specification of CMPs, NACR would identify templates of CMPs. We therefore assessed our hypothesis in the context of biological datasets.

We verify in **Figure 3** of the **Results** that NACR identifies partial templates of single-cell gene expression distributions of the human immune system and of mouse embryo development in GRCh38 and mm10 respectively. The genomic loci produced by CPSA for editing expression of the gene coding for interferon-γ (IFNG) in $CD8^+$ T cells corresponded to genes that had been confirmed in an independent genome-wide CRISPR activation screen to edit interferon-γ protein expression in $CD8^+$ T cells. The genomic loci produced by CPSA for editing $CD4^+$ T cell "effectorness" genes (Cano-Gamez et al., 2020) corresponded to key $CD4^+$ T cell regulatory modules. The genomic loci produced by CPSA for editing the composition of the embryo with respect to developmental lineages corresponded to key developmental transcription factors and identified master regulators specific to the lineages in question.

We verify in **Figure 4** of the **Results** that NACR identifies partial templates of the high-parameter tissue organization of the colorectal colorectal cancer (CRC) immune tumor microenvironment (iTME) in GRCh38. The genomic loci produced by CPSA for editing the frequency of tumor infiltrating $CD8^+$ T cells corresponded to genes, including CXCL11, as well as

the promoter region of CD62L, whose experimental genetic manipulation has been shown to modulate tumor infiltration of $CD8^+$ T cells. Furthermore, genomic loci produced by CPSA selectively highlighted an active site in CXCL11 whose mutation was shown to alter T cell trafficking in vivo by altering ligand binding. The genomic loci produced by CPSA for altering the CRC iTME from a diffuse to an organized B cell state corresponded to genes, including CXCR4, whose chemical modulation in CRC patients was shown to modulate B cell immunogenicity of the iTME. In addition, CPSA highlighted the zinc-transporter SLC30A1, whose expression was associated with survival of CRC patients and suggested a potential mechanistic role for zinc ion homeostasis in B cell mediated antitumoral immunity.

Our results, showing that PSAs of CMPs from three fundamental biological contexts (the immune system, development and cancer) are partially templates, provide validation of our theory.

Results

A **Glossary** is provided in the **Methods** as a reference for the new terms introduced.

Defining templates of complex molecular phenotypes with causal phenotype sequence alignment

We define a template to be a structure, according to which a system produces an observed structure that is formally a copy of the template. Motivating biological examples of templates are the coding sequence of a protein within the genome and the DNA encoding of a transcribed RNA. What would a template of a complex molecular phenotype (CMP, like the organization of a biological tissue or a single-cell gene-expression distribution, look like in a reference genome?

The formal structure of a CMP can be regarded as a collection of points in a metric space (a set with a notion of distance between points in the set). For example, a CMP could be: a collection of expression vectors of single-cells (Samusik et al., 2016) or a collection of cell-type composition vectors of patches from a tissue (Schürch et al., 2020). The distances between points in the metric spaces measure how different the gene expression or cell-type composition vectors are.

The formal structure of a genome is traditionally regarded as a sequence of nucleotides. We reformulated the genome as a metric space, so as to axiomatically define a phenotype sequence alignment (PSA) of a CMP with a reference genome as an exact copy of the CMP's formal structure within the reference genome: a PSA of a CMP with a genome is a map sending points in the CMP to points in the genome that preserves distances between points. We first use this axiomatic definition of PSA to define templates, before detailing how we formulate the genome sequence as a metric space.

In **Figure 1B.1** the points within the oval (left) represent the CMP and the collection of points (right) represent the genome (formulated as a metric space). The distance between the points in **Figure 1B.1** are their distances on the page. In **Figure 1B.2**, the PSA of the CMP with the genome is highlighted in red and can be seen to capture the formal structure of the CMP. Three example points, and the locations to which they are sent in the genome, are shown.

If a PSA is a template of the CMP, making a prescribed edit to a CMP should only require editing the locations corresponding to it in the PSA. We consider an edit to a CMP as moving its points (**Figure 1B.3, blue arrows indicate how selected points in phenotype space are moved, Figure 1A.3, achieving selected edit on top row to CMP requires only correspondingly editing the template in genome on bottom row**). Such an edit could correspond to, for example, editing the gene-expression of a cell type or the cell type composition of a cellular neighborhood in a tissue). Since the PSA sends each point in the CMP to points in the genome, there are genomic loci in the PSA corresponding to the edit (**Figure 1B.4, highlighted red points**). If experimental

manipulation of these genomic loci results in the organism producing the correctly edited CMP, the PSA is defined to be a template.

The procedure of computing a PSA of a CMP with a reference genome, and mapping prescribed edits of the CMP to their corresponding genomic loci in the PSA, is referred to as causal phenotype sequence alignment (CPSA). Whether a PSA is a template for a CMP is evaluated empirically by verifying whether manipulation of the genomic loci output by CPSA results in production of the edited CMP.

The notion of alignment and template we provide here is formally analogous to the notion of alignment and template for protein sequences. This is discussed in **Supplementary Note 2** using ideas from model theory (Grädel et al., 2007).

The genome as a metric space

We axiomatically defined PSAs and templates assuming that the genome had been formulated as a metric space. We now present our formulation of the genome as a metric space. We describe the points in this metric space and the distances between pairs of points. An informal description is provided here; the formal definition is in **Supplementary Note 1, Section 1.2.**

The points in the metric space representation of a genome are k-mer functionals. A k-mer functional is axiomatically a function that assigns a numerical value to every unique k-mer in the genome (**Figure 1D, left, λ_1 and λ_2 are two examples of k-mer functionals**). A k-mer functional should be visualized as a "trace" along the genome: it assigns a numerical value to each position in the genome dependent only on the k nucleotides beginning at that position (**Figure 1D, top right, red line labelled λ_1 and green line labelled λ_2**).

The distances between points in the metric space are as follows. The squared distances between two k-mer functionals is obtained by averaging the squared difference in their values

at the k-mer across each position in the genome (**Figure 1D, right, orange lines between red and green curves indicate pointwise differences**).

In practice, we cannot evaluate a functional at every genome position, because the genome is too large. Instead, we estimate the functional's values by sampling k-mers from the genome, and averaging the squared distances between functional values over the sampled k-mers (the scale factor has been excluded for clarity from the formula in **Figure 1D, lower right**; details are provided in **Supplementary Note 1, Section 1**).

A single genome is represented by the metric space of k-mer functionals, and not to an individual k-mer functional. The sequence of the genome is captured by its metric space representation: the metric on the space captures all the frequencies of k-mers. The metric space is a natural way to represent a single genome, because the primary output of short-read sequencing technologies is a distribution over k-mers (Huang et al., 2012).

Alignments of CMPs with a genome sequence

Having formulated the genome as a metric space, we now describe and visualize PSAs of a CMP with a reference genome. The mathematical definition is in **Supplementary Note 1, Section 1.3.**

A CMP has three pieces of data: 1) a set of possible, high-dimensional, measurement outcome vectors; each vector a single possible observation (**Figure 1C, left, x_1 and x_2 are vectors, assigning a value to each feature, with examples: gene expression of a single cell and cell composition of a tissue image patch**), 2) a notion of distance between the vectors (**Figure 1C, top right, formula for Euclidean distance**) and 3) the observed measurement outcome vectors in the dataset (**Figure 1C, lower right, x_1 , x_2 , … are observed**). The first two ingredients are referred to together as 'phenotype space' and the third as the 'points' of the CMP.

A PSA of a CMP with a reference genome is a function T that maps points in phenotype space to k-mer functionals (**Figure 1E, left, point x_i is mapped to $T(x_i)$**) such that: for all points x_i and x_j in

the CMP, the distance between x_i and x_j is equal to the distance between the k-mer functionals assigned to them, $T(x_i)$ and $T(x_j)$ (**Figure 1E, left, arrows between $T(x_i)$ and $T(x_j)$ are the same length as the arrows between x_i and x_j; see lower formula**).

A PSA of a CMP is visualized in **Figure 1B.4.** Concretely, the map T assigns a trace along the genome to each point in the CMP (**Figure 1E, red, green and blue traces**). The average of the squared differences in values at each position in the genome (**Figure 1E, vertical dashed lines and areas between pieces of traces indicated with colors and lowercase letters**) between the traces assigned to a pair of points has to equal the squared distances between the points.

Mapping points in phenotype space to k-mer functionals means the genomic loci corresponding to a point in phenotype space are not necessarily localized in one contiguous sequence, but can be distributed across the genome. This is as would be biologically expected, if the PSA were to be considered a template.

A neural algorithm for identifying PSAs of CMPs with reference genome

We defined a PSA of a CMP with a reference genome as a distance-preserving map, labelled T, sending points in phenotype space to k-mer functionals. Requiring that T preserves distances introduces many constraints on T (**Figure 1E, right, colored areas between colored traces have to satisfy equations**). These constraints provide a way to computationally search for PSAs of a CMP within a reference genome that does not require any biological annotations, external data or genetic variation.

We formulated the problem of assigning k-mer functionals in the genome to points in phenotype space while preserving distances as an optimal transport problem (i.e., Gromov-Wasserstein distance minimization (Mémoli, 2017; Nitzan et al., 2019)) that can be approximately solved with neural networks. We used a neural network strategy because neural networks can extract features and translate between data modalities without supervision in

biological (Yang et al., 2021) and other domains (Radford et al., 2021), as well as predict phenotype from genotype (Vaishnav et al., 2022; Zhou and Troyanskaya, 2015) and generate biological sequences achieving phenotypic targets (Vaishnav et al., 2022). Our algorithm is termed "Neural Alignment of CMPs with a Reference genome" (NACR, **Figure 2A-B, described presently and in Supplementary Note 1, Section 2**).

NACR has two inputs: a CMP and reference genome sequence (**Figure 2A, left, upper**). The output is an approximate PSA, a map T that sends each point in phenotype space to a k-mer functional (i.e., point in the reference genome represented as a metric space) that preserves distances. The map T is found by imposing that it approximately preserves distances (**Figure 2A, left, lower**) while restricting its computational complexity.

The map T is parameterized by two neural networks F and G (**Figure 2A, right, boxes labelled F and G**). F takes as input a data point x and outputs an intermediate representation vector F(x), and G takes as input a k-mer y and outputs an intermediate representation vector G(y) (**Figure 2A, right, input and output to boxes labelled F and G**). The value of the k-mer functional T(x) on the k-mer y is the inner product of their representation vectors; i.e., $T(x)(s) = \langle F(x), G(y) \rangle$.

The distance between two k-mer functionals is obtained using a Monte-Carlo estimate: by sampling a batch of sequences and averaging the squared differences between the two functionals over the batch. The extent to which T deviates from preserving distances is used as an 'alignment loss' to be minimized by gradient descent (see **Supplementary Note 1, Section 2**).

NACR is distinct from standard approaches for translating or learning joint representations of multimodal data (Radford et al., 2021; Yang et al., 2021), because 1) there is no assumed coupling between k-mers and data points 2) we are searching for a k-mer functional to assign to each point in phenotype space and not an individual k-mer.

NACR is applied to high-parameter imaging data as follows. First, the CMP – a tissue image dataset – is represented as a matrix of the counts of different cell types (after log (1+x) transformation) within image patches. The cell types are identified by segmentation and clustering as in (Bhate et al., 2022; Schürch et al., 2020) **(Figure 2B, upper)**, but no biological information about the cell types is retained**.** The distance in phenotype space is the standard Euclidean distance. Other, lower-level features of the patches (e.g. pixel intensities) could also be used for NACR. However, since we empirically evaluate whether the output PSAs are templates by editing the CMP with respect to these features and applying CPSA to the edits, we require the features to be simple and biologically interpretable.

The genome sequence is represented by extracting 128-mers (**Figure 2B, middle**). This size of k-mer was selected to enable at least 10e7 128-mers to be stored on GPU along with the single-cell distance matrices during the subsequent computations (see **Methods**).

The PSA (**Figure 2A, right**) is obtained by minimizing the alignment loss using gradient descent. The alignment loss is computed by sampling batches of image patches and k-mers, applying F and G, and measuring the difference in pairwise distances between pairs of patches in the original phenotype space and the estimated distances between their assigned 128-mer functionals in the reference genome (see **Supplementary Note 1, Section 2**).

PSAs of CMPs require complex genome sequences

If NACR produced PSAs that were templates for the input CMPs, whether or not a PSA can be identified for a given CMP should depend on the target genome sequence. If there are no constraints on how points in phenotype space are mapped to the k-mer functionals, it is possible to perfectly identify perfect PSAs of any data with any genome that has sufficiently many unique k-mer sequences. We therefore reasoned that the quality of the best PSA for a human CMP would be considerably better in GRCh38 than the reference genomes of other

species, but only when the permitted complexity of the PSAs output by NACR had been constrained.

We selected a human CMP – a 46 antibody-parameter CODEX image of a human tonsil (Kennedy-Darling et al., 2021) and applied NACR to search for PSAs with the genomes of *H. sapiens, M. musculus, E. coli, S. cerevisiae* as well as with genomes randomly generated by uniformly sampling nucleotides, while varying the permitted complexity of the PSAs. The complexity of PSAs output by NACR was varied by adjusting the regularization strengths of its neural networks (keeping the same architecture).

In detail, we applied PSA as follows. We represented the multiplexed imaging data as a matrix in which rows corresponded to image patches within the tonsil, and columns correspond to the frequencies of different cell types within patches. We next extracted (at random) 10000 128-mers (k= 128) from the genomes of *H. sapiens* (human), *M. musculus* (mouse), *S. cerevisiae* (yeast), *E. coli* as well as 10000 uniformly distributed 128-mers. We applied NACR using 1-layer neural networks (see **Methods**) with each extracted set of 128-mers. This extraction and subsequent application of NACR was repeated multiple times to account for variation in the samples of 128-mers, at multiple regularization strengths to vary the permitted complexity of the alignments (**Figure 2C, each point indicates a trained model at regularization indicated by x-axis position and is colored by the species, see Methods**).

This analysis showed that parsimonious PSAs (i.e., those of restricted complexity) can only be found in complex genome sequences. When no complexity constraint on the alignment was imposed, all genomes, including the randomly generated one, had a PSA with the tonsil data with the same, low alignment loss (**Figure 2C**, **regularization parameter 0**). Imposing modest complexity constraints, PSAs with the randomly generated genome could not be found (**Figure 2C, model regularization = 10, blue points),** whereas PSAs with the other genomes could. At a higher regularization, PSAs could be identified with the yeast, human and mouse genomes but not in the *E. coli* genome (**Figure 2C, model regularization = 100, orange points)**. With yet

stronger regularization, better PSAs could be found with the human and mouse genomes than with the yeast genome (**Figure 2C, regularization = 200**). At the strongest regularizations, better PSAs were found within the human genome than in the others (**Figure 2C, regularization parameter 500-1000**).

These data indicate that the reference genome sequence affects the quality of the PSAs identified by NACR and suggest that the human tonsil could have the best PSAs with the human genome.

PSAs capture genetic information specific to CMPs

If the PSAs identified by NACR were templates, the PSAs should capture genetic information biologically specific to the input CMPs. We compared the known genetics of biologically defined points in the phenotype space to the k-mer functionals assigned to those points in the PSA.

A PSA sends any point in phenotype space to a k-mer functional in the genome. A PSA therefore highlights regions of the reference genome sequence, namely those where the k-mer functional takes high values. If the k-mer functionals corresponding to specific, biologically defined, points in phenotype space highlights regions of the genome involved in the genetic regulation of the biological processes associated with those points – even though no biological annotations or external data were used to find the PSA – it could be concluded that the PSA was capturing genetic information specific to the CMPs.

We found a PSA of the human tonsil CMP with the human genome (GRCh38). Using the same patch-feature representation of the human tonsil in terms of cell type counts as above (**Figure 2B, upper panel**), we scaled up PSA. Specifically, we extracted ten million 128-mers (a coverage of approximately 0.37, using a haploid genome size of 3.1e9 base pairs), increased the capacity of the neural networks and imposed nonnegativity (for interpretability) as well as a sparsity

regularization term in the loss (to encourage the functionals to be nonzero at only a small number of 128-mers, see **Methods**).

We next selected two points in phenotype space of known biological significance: the mean cell type count vector of points in the light-zone of the germinal center, and the mean cell type count vector of points in the dark-zone of the germinal center, using the cellular neighborhoods defined in (Bhate et al., 2022). These points in phenotype space represent the average cell-type composition of image patches randomly sampled from the dark-zone and light-zone respectively. We selected these points because the genes involved in regulation of B cell affinity maturation via antigen-dependent, T cell mediated activation in the light-zone followed by movement to the dark-zone and rapid proliferation or apoptosis (Ise and Kurosaki, 2019) are well understood.

We evaluated whether the k-mer functionals assigned in the PSA to the dark-zone cell-type count vector and light-zone cell-type count vector highlighted the genes associated to those cellular neighborhoods in the tonsil.

From the sequence of each gene in the human genome, we extracted 128-mers in the region between its transcription start site (TSS) and transcription end site and computed the average difference between the functionals T(x) and T(y) assigned to the average light-zone cell type frequency (x) and dark-zone cell type frequency (y) respectively to obtain a "gene functional", T(x)-T(y) (gene). The value of this functional on a gene is used for downstream gene-set enrichment analysis (**Figure 2D, k-mer function values on each gene**).

Ranking genes by the gene functional T(x)–T(y), we performed gene-set enrichment analysis (GSEA) (Korotkevich et al., 2021) with Gene Ontology Biological Process gene sets (Ashburner et al., 2000). 19 biological process gene sets were differentially highlighted by the functionals corresponding to the light zone and the dark zone (FDR = 0.1) (**Figure 2E, table and Supplementary Tables, HLT GSEA**). These gene sets included processes specific to the immune

response, including apoptosis associated genes enriched by the dark-zone functional, signaling-modulation genes enriched by the light-zone functional, as well as lymphocyte chemotaxis and humoral immune response genes. Other gene sets that were significantly different between the functionals included response to zinc ion and cadmium ion, potentially suggesting a role for these pathways in the regulation of the germinal center response.

When this analysis was performed multiple times by randomly initializing but not training the neural networks in NACR, these gene-sets were not significantly enriched, but others were (**Figure 2E, table and Supplementary Tables, HLT GSEA, columns labelled 'train FDR'**). The PSA obtained by randomly initializing but not training the neural networks of NACR is referred to as a "null alignment". Significantly enriched gene sets in null alignments could be explained by the fact that some gene sets contain genes with a high degree of conserved sequence: if by chance a certain k-mer obtains a high value from a functional, then the other genes in that gene set are likely to obtain a high value from that functional; gene sets with high sequence similarity are more likely to be significantly enriched.

Was the observed genetic specificity of the PSA identified by NACR specific to the human tonsil CMP? We performed the same analysis with a CMP from a different biological context, the colorectal cancer (CRC) immune-tumor microenvironment (iTME). If NACR identified a PSA in this context that was biologically specific to CRC, it would suggest that the specificity observed for the human tonsil multiplexed imaging data was not a coincidence of the statistics of the sequence similarity of the gene sets involved in the germinal center response, or some other artefact of our sequence extraction procedure.

We applied NACR to align a 56 antibody-parameter CODEX dataset of iTME samples from CRC patients (Schürch et al., 2020) with GRCh38 and analyzed two points in phenotype space: the average cell type count vectors for patches from the two patient groups in that dataset (instead of the light-zone and dark-zone of the germinal center as with the tonsil above). The distinction between patient groups in this dataset is that in one (termed Crohn's like reaction, CLR), the

cancer is less aggressive and tertiary lymphoid structures (TLS) are formed, whereas in the other (termed diffuse immune infiltrate, DII), they are not; this as well as other differences in tissue organization are associated with survival outcomes (Schürch et al., 2020).

We aggregated the difference in functionals assigned to CLR and DII patients' average cell type count vectors in the PSA over genes and performed GSEA. 30 biological process gene sets were enriched (**Figure 2F and Supplementary Tables CRC GSEA**) after removing those appearing in null alignments (FDR < 0.1). These included the morphogenesis of a branching epithelium (CRC is a cancer of epithelial cells), wnt signalling (a key pathway in the development of colorectal cancer) and vasculogenesis and extracellular matrix organization (crucial for defining the architecture of the iTME) (Markowitz and Bertagnolli, 2009). The other gene sets corresponded to developmental processes (such as retinoic acid response and embryonic development), which is consistent with the hypothesis (Fessler and Medema, 2016) that one patient group's tumors are driven by a hijacking of developmental processes.

Thus, in two CMPs from distinct phenotypic contexts (the human tonsil and colorectal cancer), the PSAs identified by NACR captured specific genetic information to those phenotypes, even though no biological annotations or external data had been utilized to find the PSAs.

PSAs are associated with human genetic variation

Given that the difference between functionals assigned in the PSA to the CLR and DII mean cell-type frequency vectors highlighted specific gene modules involved in CRC, we assessed whether this functional was associated with human genetic variation implicated in colorectal pathology. We obtained from ClinVar (Landrum et al., 2018) 419 pathogenic single-nucleotide polymorphisms (SNPs) that were labelled as expert validated and pathological, contained the search-term "colorectal" and did not contain indeterminate (N) nucleotides in the 64 nucleotides either side in GRCh38 (so we could compute the functional at the 128-mer centered at the SNP) (**Supplementary Tables, colorectal SNPs)**. We found that the average

values of the functional on the 128-mers centered at these SNPs were significantly higher (permutation p = 0.004) than on random 128-mers extracted from the 1Mb regions either side of each SNP (**Figure 2G, values of functional on SNPs and non-SNPs individually**). Only 3% of null alignments obtained the same or smaller p-value, indicating that the difference can only partially be explained by NACR biasing values of particular sequences to be correlated. These data further corroborate that the PSAs identified by NACR capture genetic information specific to input CMPs.

PSAs are not proxies for gene expression signatures

One possible explanation for why functionals in the PSAs captured specific genetic information without incorporation of any biological annotations or genetic variation could be that the PSAs output by NACR somehow highlighted sequences that were proxies for local gene expression signatures associated with the tissue image patches. If so, it would be expected that functionals would preferentially highlight genic regions relative to other regulatory regions, and that this would be similar across PSAs for different CMPs.

We randomly extracted 10000 128-mers from transcript, promoter (defined as 2kb upstream of a transcript), terminator (2kb downstream) and intergenic (not transcript, promoter or terminator) regions of each chromosome, and evaluated the absolute values of the functionals in the PSAs of the tonsil and CRC iTME analyzed above (**Figure 2H, described presently**).

The functionals for both the CRC and tonsil PSAs took a significantly higher absolute value on promoter, transcript and terminator regions than intergenic regions (**Figure 2H, permutation p-values indicated in Supplementary Figure 1A**). In addition, the absolute values of the functionals were overall significantly higher for promoter and terminator regions than for the transcript regions. In the case of the CRC patient group difference functional, this was consistent across chromosomes (**Figure 2H, Supplementary Figure 1A**). However, the tonsil dark-zone/light zone functional took a significantly higher value on transcript regions in

chromosomes 19 and 20 relative to the other regions. Thus, the identified PSAs cannot be viewed entirely as identifying gene expression signatures associated with each point in phenotype space, and in some cases preferentially learn signatures associated with regulatory regions. The differences between the tonsil and CRC functionals indicate that the value differences are not merely a property of the genome's composition in terms of the promoter, terminator and transcript regions.

NACR produced PSAs of two CMPs (tissue organization of human tonsil as well as of CRC) that highlighted the genes involved in the regulation of those CMPs, with no *a priori* biological annotations and only the reference genome and proteomic imaging data as input. This suggests that the PSAs identified by NACR capture genetic information specific to the input CMPs. In addition, since null alignments did not yield significant results, this observed specificity is not an artefact of the inductive biases imposed by the network, or of our sequence extraction and evaluation methodology.

Implementing causal phenotype sequence alignment

Aligning CMPs with the human genome using PSA captured genetic information specific to the CMPs. Could the resultant PSAs be templates? We developed "causal phenotype sequence alignment" (CPSA), which produces predictions whose empirical validation justify regarding a PSA as a template.

Given an PSA of a CMP with a reference genome, CPSA maps a CMP edit to the genomic loci in the PSA that correspond to the edit (recall in **Figure 1B.4** how the PSA highlights the red points in the genome corresponding to the edit). If the PSA is a template, experimentally manipulating the genome at the loci output by CPSA should result in production of the edited phenotype.

The input to CPSA is a PSA (e.g., produced by NACR) and an edit to the CMP. The output is a k-mer functional referred to as a "discrepancy functional" for the edit. A PSA is a template if,

given an edit to the CMP, experimental manipulation of the genomic loci where the discrepancy functional for the edit is high, is sufficient to achieve the edit. The mathematical definitions and computational methodology of CPSA are provided in **Supplementary Note 1, Section 3** and **Methods,** and are informally described below.

The first step in defining CPSA is to specify an edit to a CMP. A standard way to represent a CMP is to cluster its points (e.g. clusters of cells by gene expression are cell states and clusters of image patches by cell type composition are cellular neighborhoods). The green regions in **Figure 3A, top left oval** are clusters x_i with sizes corresponding to the number of observations assigned to the cluster. For our first experiments, the edits we considered were small shifts in the means of clusters (**Figure 3A, top left oval, shifting the mean of cluster 1 by dx_1 as indicated by blue arrow results in the new cluster in red**). Such edits correspond to, for example, a change in the gene expression of a cell type or change in the cell type composition of a certain cellular neighborhood.

The PSA assigns a k-mer functional to the original cluster mean and to the new cluster mean (**Figure 3A, top right traces, red and green functionals assigned by PSA to $x_1 + dx_1$ and x_1 respectively**). At each k-mer, the squared difference between the functionals assigned to the new and original cluster means measures how much the original and edited CMP differ at that k-mer: a "discrepancy" between the original and target CMP (**Figure 3A, blue bars between red and green traces)**. Since the discrepancy is a value at each k-mer, we obtain a k-mer functional (**Figure 3A, bottom left, formula for dT_{edit}**), and visualize it as a trace (**Figure 2A, blue bars between red and green traces are plotted at each position in the genome in trace below)**.

We define a PSA to be a template when, given an edit to a CMP, manipulating the genomic loci with high discrepancy for the edit output by CPSA (**Figure 2A, highlighted region on the genome sequence where discrepancy is high**) results in production of the edited CMP.

A PSA being a template can be experimentally evaluated by assessing whether manipulating the genomic loci with high discrepancy for a given phenotypic edit achieves the edit. Manipulation can be either by directly mutating with genome editing, or altering their downstream influence with sequence specific transcriptional modulators or chemical inhibitors.

Information-efficiency of the genome enforces CMPs to have simple templates

Which evolutionary pressures could explain why CMPs could have templates in the genome, and justify searching for templates of CMPs from biological datasets? We reasoned that selective pressure on the genome to specify CMPs efficiently (in an information theoretic sense) could ensure that they have templates. We devised a simulation strategy to assess whether PSAs would be templates under such constraints (**Figure 3B, described presently and in detail in the Methods**).

In our simulations, we randomly generate a genome sequence and a genotype-phenotype map (which outputs a CMP from a genome) in the presence or absence of information-efficiency constraints. Knowing the genotype-phenotype map means an output CMP can be obtained from a mutated genome. We can therefore assess whether the CMP has templates, as follows:

1. Obtain a PSA by applying NACR applied to the original CMP.
2. Obtain a discrepancy functional by applying CPSA to a prescribed phenotypic edit using the PSA from step 1.
3. Mutate the genome at the genomic loci with high discrepancy.
4. Apply the known genotype-phenotype map to the mutated genome to obtain a new CMP.
5. Assess whether the new CMP has the correct edit

If the loci with high discrepancy are those whose mutation results in the new CMP being close to the correctly edited CMP, the PSA is a template.

In detail, the simulation was conducted as follows. We began by sampling a random binary genome of 1000 positions (**Figure 3B, upper, binary genome g of length n = 1000**). Next, we trained a deep neural network (intended to model the "true" genotype-phenotype) to have as output a CMP that was a mixture of two Gaussian distributions, with means 1 and -1 (**Figure 3B, upper, neural network Φ with Gaussian mixture output**). The network had just one input-output pair as training data (the random string and the output). Information efficiency constraints on the genotype-phenotype map could be enforced by imposing regularization on the weights of the network as it was trained.

Next, NACR finds a PSA T of the CMP output by the genotype-phenotype map with the input genome (using 3-mer functionals) (**Figure 3B, lower, left, top**). We considered the edit to be a shift in one of the means of the Gaussian mixture, from 1 to 1.5 (**Figure 3B, lower, left, blue histogram indicates target shift**), a shift in cluster means as in **Figure 3A**. Therefore, CPSA outputs a discrepancy functional, $(T(1) - T(1.5))^2$ (**Figure 3B, lower, middle**).

We can change the genome at each position (from 1 to 0 or vice versa) and apply Φ (**Figure 3B, lower, right, blue indicates target phenotype and orange indicates phenotype obtained by mutating the genome and applying Φ**). We measure the correlation between the discrepancy at each genomic position and the distance to target phenotype obtained by changing the observed sequence at that position (from 1 to 0 or vice versa) and applying the known genotype-phenotype map Φ (**Figures 3C, each line in each violin indicates one of twenty starting genomes to which the simulation strategy was applied and the median correlation between discrepancies obtained by twenty random initializations of phenotype-genotype embedding).** If the PSA is a template, there should be a negative correlation between discrepancy and distance to target phenotype.

The simulation confirmed that there was a significant, negative correlation (in relation to null alignments) between the discrepancy at a position (averaged over 3-mers at that position) and the distance to the target phenotype. Null alignments had positive correlations, and the

magnitude of the negative correlation was significantly greater in trained templates than in untrained templates (**Figure 3C, compare absolute values of blue and orange violins above unregularized to those of untrained respectively**). The difference in correlation (between PSAs and null alignments) was greater in the presence of information-efficiency constraints (**Figure 3C, compare difference in absolute value of blue, unregularized and blue, untrained violins to difference in absolute value of orange, unregularized and blue, untrained, p-values indicated by * and ** respectively).**

Thus, if the genome had to specify CMPs under the presence of information-efficiency constraints, CPSA would produce discrepancies for an edit that reveal loci whose manipulation would result in production of the correctly edit CMP. That is, NACR would identify PSAs of CMPs with the reference genome that were templates.

CPSA output is consistent across PSAs obtained from biological replicates

Simulations showed that NACR would identify templates of CMPs. This result justified assessing whether NACR identified templates in biological datasets. We evaluated this in the context of single-cell gene expression (**Figures 3D-I**) and subsequently tissue organization (**Figure 4**).

The edits to single-cell gene expression distribution CMPs we consider are shifting the gene expression in a cell type. A discrepancy functional is obtained as follows. First, NACR is applied to a single-cell gene expression matrix to obtain an PSA, T (**Figure 3D, upper**). Next, CPSA outputs a discrepancy functional for shifting the mean gene expression $\mathbf{x_{g,c}}$ of a gene **g** in a cell type **c,** as follows. The discrepancy functional is the squared difference $(T(\mathbf{x_{g,c}}') - T(\mathbf{x_{g,c}}))^2$ where $\mathbf{x_{g,c}}'$ is obtained from $\mathbf{x_{g,c}}$ by increasing the value of gene g (**Figure 3D, T(original expr.) and T(shifted expr.) respectively)**. The k-mers with a high discrepancy are identified (**Figure 3D, lower, squared difference between green and blue functionals is the discrepancy, see Methods)**.

We first confirmed that the discrepancies for shifting the mean expression of cell types were consistent across biological replicates (**Figure 3E,** described presently). We obtained a single-cell gene expression matrix of untreated peripheral blood mononuclear cells (PBMCs) in healthy donors (Hao et al., 2021). We transformed the data to 1000 PCs and used NACR (with 1-layer neural networks and varied regularization) to find PSAs of each donor's time 0, untreated data.

The edits to the CMP we considered were shifts in the mean gene expression vector of the most abundant cell types (monocytes, $CD4^+$ T cells and $CD8^+$ T cells) in the direction of each of the top three principal components (PCs). We extracted 50000 128-mers from transcript regions of GRCh38, applied CPSA for each edit to each donor's PSA with GRCh38 and evaluated the resulting discrepancy functionals on the extracted 128-mers. We then evaluated the correlations between the discrepancies for each pair of donors (**Figure 3E for select regularization and first two PCs, Supplementary Figure 3C for other regularizations and PCs, points represent pairs of donors**).

The distribution of correlations between pairs of donors resembled that of null alignments when NACR was not regularized (**Figure 3E**, **compare grey points to red points for each cell type and PC**). We observed positive correlations between pairs of donors when NACR was trained with both sparsity regularization and complexity regularization (i.e., regularization on neural network weights, see **Supplementary Note 1, Section 2**) that was consistent across cell types and PCs (**Figure 3E, compare brown and purple points**), but this was diminished by excessive regularization (**Supplementary Figure 3A, points indicate correlations at high sparsity and complexity regularization for each cell type and PC)**. Including sparsity regularization favored stronger correlations, but added variance, with pairs of donors possessing a greater number of strong negative correlations (**Figure 3E, compare purple and brown purple points**). Taken together, these findings indicate that discrepancy functionals from the PSAs produced by NACR are generally consistent across biological replicates, although the consistency is sensitive to choice of hyperparameters.

PSA of PBMC single-cell gene expression with GRCh38 is a template

If a PSA of single-cell gene expression with a reference genome is a template, the genes whose experimental manipulation successfully achieves an edit to the gene expression of a cell type in the CMP (so-called "hit-genes") should be those with a high average discrepancy across their k-mers for that edit, produced by CPSA. Moreover, this association (between hit-genes and highly discrepant genes) should be specific to shifting the target gene – discrepancies for shifting expression of other genes should not be associated with hit-genes.

We obtained a genome-wide CRISPR activation screening dataset, in which each gene had been evaluated for its ability upon activation to modulate expression of interferon-γ in human $CD8^+$ T cells derived from PBMCs (Schmidt et al., 2022). If a PSA of the PBMC gene expression data with GRCh38 were a template, the hit genes in the CRISPR screen should be those with a high discrepancy output by CPSA for increasing the mean $CD8^+$ T cell gene expression vector in the direction labelled by IFNG (the interferon-γ gene).

A perfect overlap of genes of high discrepancy from CPSA and hit genes in the CRISPR screen would not be expected for at least three reasons: (1) the CRISPR screen corresponded to transcriptional activation and not sequence manipulation, (2) the screen assessed protein expression and so includes hits operating at the translational level (3) genes whose modulation indirectly upregulated interferon-γ in $CD8^+$ T cells (e.g. by affecting other cell types) would not be observed in the CRISPR screen, but may be the discrepancies from CPSA. We applied NACR to obtain a PSA of the single-cell gene expression distribution of healthy PBMCs with GRCh38 and confirmed using CPSA that the alignment was causal (**Figures 3F-G**, described presently and see **Methods** for implementation details).

We obtained the discrepancy functionals for increasing (multiplying by two in log space) the expression of the 223 genes in the PBMC dataset that had a similar average expression and detection frequency as IFNG (including IFNG), and aggregated these discrepancies over gene

sequences (as detailed in **Figure 2D**) to compute gene-level discrepancies (**Supplementary Figure 3A, each column corresponds to a target gene to increase, and each row to a gene over which discrepancies were aggregated over 128-mers**).

The discrepancies were correlated across target genes (**Supplementary Figure 3A, columns have similar color patterns**). This could be explained by the fact that the functionals in the alignment output by PSA are nonzero at only a sparse collection of 128-mers (as a result of the regularization) and so genes with these 128-mers are likely to be enriched amongst genes with a high discrepancy with respect to any pair of functionals. Accordingly, we removed this common signal by removing the first singular value of the matrix of discrepancies (**Supplementary Figure 3B, columns no longer have as similar color patterns and see Methods**).

The median gene level discrepancy for hit genes was significantly higher (permutation $p < 0.05$) than non-hits at multiple thresholds for defining genes to be hits in the CRISPR screen, and the area-under-the receiver operator characteristic curve (AUC) obtained using high discrepancy to define hits was significantly higher (permutation $p = 0.05$) than a random classifier at multiple thresholds for defining genes to be hits in the CRISPR screen. (**Supplementary Figure 3**, **blue line corresponds to permutation p-value for difference in medians between hits and non-hits and orange for AUC of hits, at differing hit thresholds**). The best threshold AUC was 0.528, obtaining a permutation p-value of 0.02 (**Figure 3F**, **receiver operator curve**). We confirmed that this result did not depend on the choice of genes used for normalization by sub-sampling 20 genes at a time and repeating the p-value computation (**Supplementary Figure 3F, histogram of p-values**). Thus, the genomic loci identified by CPSA correspond to genes whose experimental manipulation achieves the desired phenotypic edit.

If the PSA were a template, only a small fraction of genes (including IFNG) should have the property that discrepancies for increasing them are associated with the CRISPR screen hits genes for interferon-γ expression. We computed the permutation p-value and AUC when using

the normalized gene level discrepancies for increasing the mean expression in each of the 223 genes' directions. Only 14/223 (6.2%) of other genes had a discrepancy associated with hit genes(**Figure 3G**, **black dots corresponding to AUC permutation p-value < 0.05**), and no gene had a lower AUC p-value than IFNG (**Figure 3G red star indicates IFNG**). Since all the genes had a similar mean and detection frequency as IFNG, this was not due to technical differences in gene expression patterns. Thus, the association between discrepancies and genes empirically established to modulate interferon-γ expression was specific to shifting the expression of IFNG in $CD8^+$ T cells.

One possible explanation for these results would be if the discrepancies were simply highlighting the sequences of genes correlated in expression space with IFNG in $CD8^+$ T cells. There was no association between the normalized IFNG-increase discrepancy of each gene and its correlation across $CD8^+$ T cells with IFNG (**Supplementary Figure 3E, Spearman r = 0.001, p = 0.87**). When the discrepancies were adjusted by taking residuals from a linear regression model with correlation as variable, the permutation p-values for difference in medians and AUC were <1/5000 and 0.85 respectively. These results indicate that the gene-level discrepancies capture information regarding the sequences of genetic modulators of interferon-γ expression that is not explained by simply identifying the sequences of genes with correlated expression.

Could the highly discrepant genes identified by CPSA for increasing IFNG achieve an increase in interferon-γ production in $CD8^+$ T cells in indirect ways that could not be detected by a CRISPR screen of gene overexpression in $CD8^+$ T cells? For example, such genes could affected the behavior of other cell types. The top 5 most discrepant genes for increasing IFNG expression were: TSSK2, CFL1, SDF2L1, CAMP and ZNHIT2 (**Supplementary Figure 3F**). The CAMP gene (encoding the antimicrobial peptide referred to as LL37 (Davidson et al., 2004)) is not expressed in T cells (Uhlen et al., 2019)(Monaco et al., 2019; Schmiedel et al., 2018), but T cells co-cultured in the presence of dendritic cells induced in the presence of LL37 produced significantly higher amounts of interferon-γ than T cells co-cultured with dendritic cells that had not (Figure 6 in (Davidson et al., 2004)). Thus, the discrepancies may be highlighting the

sequences of genes whose manipulation induces the target phenotypic alteration in indirect, intercellular ways.

Discrepancies for editing immune cell gene expression highlight key genetic regulators

It was not feasible to investigate whether discrepancies produced by CPSA for any possible edit to a CMP yielded the loci whose experimental manipulation achieves the correct edit. However, given that gene expression is organized into regulatory modules, systems-level evidence that an alignment of a single-cell sequencing CMP in the genome were a template would be if the discrepancies produced by CPSA for shifting gene expression of multiple genes in a given cell type highlighted the known transcriptional regulators of those genes in the given cell type.

We obtained a list of "effectorness genes" in $CD4^+$ T cells from (Cano-Gamez et al., 2020), and applied CPSA to the PSA of the PBMC dataset with GRCh38 to obtain discrepancies for increasing expression of each effectorness gene (one at a time) in $CD4^+$ T cells (**Figure 3H**, **columns indicate target genes increased, and rows indicate genes over which discrepancies are aggregated, union top 100 discrepant genes for each target genes included**).

Genes were clustered by which edits (i.e., by which effectorness genes) they had a high discrepancy for (**Figure 3H, row colors**). We assessed enriched gene sets in each of the clusters of genes using EnrichR (Kuleshov et al., 2016) and found significant enrichment of gene sets corresponding to immune regulatory modules (**Figure 3H, color labels underneath heatmap and Supplementary Tables, PBMC EnrichR**). These included clusters significantly enriched (FDR < 0.1 in each cluster) for genes corresponding to the GO categories of: cellular response to oxidative/chemical stress (**Figure 3E, green cluster**), regulation of receptor signaling via JAK/STAT (**Figure 3H, lime cluster**), positive regulation of cytokine production (**Figure 3H, purple cluster**), and regulation of programmed cell death (**Figure 3E, lilac cluster**). These Gene Ontology biological processes correspond to key processes in the gene regulation of $CD4^+$ T cell

functional states (Cano-Gamez et al., 2020). These results support the hypothesis that the alignment for the immune single-cell gene expression CMP with GRCh38 is a partially a template.

Discrepancies for editing lineage embryo lineage highlight developmental regulators

Could NACR only identify PSAs that were templates in the context shifting the mean cell type expression of human PBMCs? We generalized CPSA from assigning discrepancies to shifting cluster means to more general changes in distribution, using optimal transport (**Supplementary Figure 3G**, **both cluster means and sizes are changed in the target phenotype, see** (Peyr\'e and Cuturi, 2019) **for background on optimal transport and** (Schiebinger et al., 2019) **for applications to comparing single-cell gene expression distributions**). In this context, CPSA uses the optimal transport coupling between original and target phenotypes is used to weight the pairwise differences in functional values (**Supplementary Figure 3G, compare formula to Figure 3A**) to compute the discrepancy functional (see **Methods** and **Supplementary Note 1, Section 3**).

We obtained a PSA of single-cell gene expression of mouse organogenesis (Cao et al., 2019) with the mouse genome (mm10) using NACR and applied CPSA to compute the gene-level discrepancies for increasing the frequency of cells belonging to each of the individual developmental lineages defined in the original publication (**Figure 3I**, **columns in heatmap indicate lineages to increase, and rows correspond to gene-level discrepancies; union of top 100 genes by discrepancy for increasing each lineage shown**).

Genes were clustered by which edits (i.e., by which lineages to increase) they were assigned a high discrepancy by CPSA (**Figure 3I, row colors**), and these clusters were enriched for master regulators of different developmental lineages (**Figure 3I, color annotations present the genes and Supplementary Tables, MOCA EnrichR**). These included a cluster enriched with genes for

neuron differentiation, bone morphogenesis and epithelium morphogenesis (**Figure 3I, pink row cluster**) that had a high discrepancy for increasing the frequencies of epithelial cells, subsets of neurons as well as chondrocytes, but a lower discrepancy for increasing the frequency of neural progenitors and endothelial lineages. Another cluster was enriched for key transcription factors involved in muscle cell differentiation, neuron differentiation and artery development (**Figure 3I, grey row cluster**), and these genes had a high discrepancy for muscle and neuron lineages. Thus, the discrepancies corresponding to increasing the frequencies of specific lineages identifies distinct modules of developmental regulation.

We further investigated the most discrepant genes for certain lineages (regressing out the discrepancies for other lineages to identify the signals most specific to each lineage, see **Methods**). We found that the top 20 genes for: increasing the inhibitory neuron lineage contained Foxg1, a transcription factor specifically involved in the regulation of inhibitory neuron differentiation (Hou et al., 2020); for increasing excitatory neuron lineage contained Tbx1 (a transcription factor involved in the balance of inhibitory and excitatory neuron subtypes (Flore et al., 2017)) and contained Klf6 for the white blood cell lineage (a key transcription factor in embryonic specification of hematopoietic lineages (Matsumoto et al., 2006)).

Taken together, our results show that in the context of single-cell gene expression CMPs of mouse organogenesis and human PBMCs, NACR produces PSAs with the respective reference genomes that are, at least partially, templates.

Editing tumor infiltrating $CD8^+$ T cell frequencies with PSAs of the CRC iTME

We reasoned that tissue organization CMPs would also have templates. That is, applying CPSA to edits to tissue organization CMPs, using PSAs with the appropriate reference genomes, obtained by applying NACR to high-parameter imaging data, would highlight genomic loci whose manipulation would achieve the desired edits to the CMPs.

While cancers are genetically heterogeneous, the established presence of phenotypic archetypes and genetic drivers raises the question of whether the organization of their microenvironments could nonetheless have templates in the reference genome. The dataset of the colorectal cancer (CRC) immune tumor microenvironment (iTME)(Schürch et al., 2020) therefore provided a therapeutically significant CMP for investigating this question.

The frequency of tumor infiltrating $CD8^+$ T cells has been shown to be prognostic for survival (Bruni et al., 2020). In the CRC iTME dataset (Schürch et al., 2020), 9 cellular neighborhoods (CNs, i.e. tissue subcompartments) are identified, one of which is the "main tumor". We therefore set our target edit to the CMP as increasing the $CD8^+$ T cell frequency within the main tumor CN (**Figure 4A, increasing the black cells in the green region**). Any genetic regulators of this phenotype could suggest therapeutic strategies. Given an PSA of the iTME with GRCh38, the discrepancy functional of interest was therefore the difference between the functional assigned by the alignment to the average cell type count vector of patches assigned to the tumor CN, and the functional it assigned to that vector shifted in the $CD8^+$ T cell direction (here, a log fold change of 1.5 was applied to shift the original cell type counts).

We first confirmed that discrepancies were consistent across PSAs obtained from multiple random initializations of the neural networks in NACR. We trained NACR six times from different random initializations to identify six PSAs of the CRC iTME dataset with GRCh38. We obtained gene-level discrepancies for the target edit by applying CPSA to these six PSAs as well as to 100 null alignments (i.e., obtained by initializing but not training NACR). Considerable heterogeneity was observed, in terms of the correlations between between gene-level discrepancies obtained from the PSAs (**Figure 4B, violin on right side**). However, there was a significantly higher correlation (permutation p-value = 0.0195) amongst pairs of PSAs than pairs of null alignments (**Figure 4B, difference between violins**).

We analyzed the gene-level discrepancies for the target phenotypic edit. For each gene, we assessed the difference between the average discrepancy (z-normalized across genes) obtained

for that gene by applying CPSA to the PSAs, versus those obtained by applying CPSA to the null alignments. We evaluated the top 20 genes with increased discrepancy in the PSAs by FDR relative to null alignments (**Figure 4B**, **points indicate genes, x axis indicates t-statistic**).

The gene CXCL11 had a significantly (FDR < 0.1) higher discrepancy in PSAs relative to null alignments. CXCL11 has been empirically demonstrated in recent studies to be a key modulator of intratumoral $CD8^+$ T cell frequency and function in multiple cancer models (Moon et al., 2018; Vollmer et al., 2021) and to be correlated with $CD8^+$ T cell frequency and survival in colon cancer (Cao et al., 2021).

We found that 15 of the top 20 genes (ranked by increase in average discrepancy from PSAs vs. null alignments) had significant associations with the expression of the CD8A gene in the Cancer Genome Atlas (TCGA) colorectal cancer (COAD-READ) dataset (Liu et al., 2018) - a proxy for intratumoral CD8 frequency in colorectal cancer (**Figure 4B, y axis corresponds to Spearman correlation with CD8A, circles indicate genes with significant associations, Bonferroni p<0.05**). When the same analysis was performed with null alignments (taking a collection of six null alignments, selecting top 20 by t-score for increase relative to other null alignments), only 3.7% ($p = 0.037$) of such models had 15 or more genes significantly associated with CD8A expression. As another statistical test, 53% of genes had a significant correlation with CD8A expression, so a random subset of 20 genes has fifteen or more genes with a significant correlation only 3.4% of the time (binomial p-value = 0.034).

Thus, the discrepancy associated with increasing the average $CD8^+$ T cell frequency in the tumor CN in trained models selectively identifies genes associated with intratumoral $CD8^+$ T cell frequency, as well as one (CXCL11) that is empirically proven to modulate the desired frequency in related tumor models.

Discrepancies for editing intratumoral $CD8^+$ T cells highlights a CXCL11 active site

Could the discrepancies from CPSA for the target edit reveal genomic loci modulating the target phenotype at the level of protein biochemistry? We tiled the CXCL11 transcript with overlapping 128-mers and assessed which positions had a significantly higher normalized discrepancy for increasing $CD8^+$ T cells in PSAs relative to null alignments (**Figure 4D, x position of each point indicates the center of the 128-mer and y position the p-value**). By far the greatest difference was found at a position within the coding sequence in the third exon (**Figure 4D, bold point between third green line and third red line and between orange lines**). We translated this and the surrounding coding sequence to amino acids and saw that the site with high discrepancy corresponded to K59 (using the position numbering in (Severin et al., 2010)). Mutagenesis of this residue was shown in (Severin et al., 2010) to modulate T cell trafficking in a peritoneal recruitment assay via alteration of biochemical protein/ligand interactions. Thus, the discrepant 128-mers identified by CPSA cannot be viewed as proxies for differential gene expression, since it also selectively identifies this protein residue that modulates the target biological function by altering protein function.

Might the discrepancies from individual PSAs identify sequences of genes that are not captured when averaging over multiple PSAs? We further investigated the discrepancies assigned by the alignment in **Figure 2**, since it had shown evidence of capturing CRC-specific genetic information.

We analyzed the top 100 most discrepant genes for increasing intratumoral $CD8^+$ T cells, and found a significant enrichment (FDR = 0.03) for two biological process gene ontology terms: killing by host of symbiont cells (ROMO1, DEFA5, CFHR1) and regulation of neuroinflammatory response (CD200R1, MMP3, PTGS2) (**Figure 4E, table**). The three genes CD200R1, MMP3 or PTGS2 in the TCGA colorectal cancer cohort were associated ($p = 0.05$) with overall survival in colorectal cancer (Gao et al., 2013) (**Figure 4D**) and CD200 and MMPs are current immunotherapeutic targets in multiple cancers (Liang et al., 2021; Xiong et al., 2020). In addition, Romo1 overexpression in adoptively transferred bone-marrow cells led to an increased intratumoral T cell frequency in a mouse model of glioblastoma (**Figure 4G, adapted**

from Figure 2 in (Sun et al., 2020, p. 1) **under the CCBY 4.0 license**). In addition, Ptgs2 knockout led to an increased intratumoral CD8 T cell frequency in a mouse model of pancreatic adenocarcinoma (**Figure 4H, right, adapted from Figure 6 in** (Markosyan et al., 2019) **under the CCBY4.0 license**). Although these validations were in tumor models distinct from colorectal cancer, the results indicate that discrepancies from CPSA could highlight genetic modulators of the desired phenotypic edit.

The analysis of functional values across different types of genomic regions (**Figure 2H)** suggested that biological information may be encoded in functional values over un-transcribed regions. We therefore aggregated the discrepancy functional for the target edit over promoter sequences of genes (5kb upstream of the transcription start site) discrepancy for the target phenotype (increasing intratumoral $CD8^+$ T cell frequency) and assessed the enrichment of gene-ontology categories in the top 100 most discrepant promoters. The only Gene Ontology category that was significantly enriched amongst these corresponded to 'leukocyte tethering and rolling' (**Figure 4E, table**), and this was due to a high discrepancy assigned by the model to the selectin (SELL, SELE and SELP) promoters. Genetic engineering of L-selectin expression by alteration of its promoter and including a proteolysis resistant domain was sufficient to increase intratumoral $CD8^+$ T cell infiltration in a subcutaneous murine tumor model (Watson et al., 2019). Thus, the identifies discrepancies may, in addition to transcript sequences and individual protein active sites, highlight regulatory regions whose manipulation can achieve the target phenotype.

Editing immunogenicity of the CRC iTME with CPSA

The CRC dataset consists of two patient groups, termed "Crohn's like reaction" (CLR) and "diffuse immune infiltrate" (DII). The CLR and DII patient groups are characterized by the presence of B cell tertiary lymphoid structures (TLS, **Figure 4J, TLS on left in yellow**) and the CLR patient group has significantly better survival outcomes (Schürch et al., 2020); altering the DII patients' iTMEs to resemble the CLR patients' ones could therefore present an important

potential immunotherapeutic strategy. We therefore applied CPSA to find the discrepancies for this particular edit to tissue organization.

We represented the original phenotype as a matrix in which each row corresponded to the average cell type composition of a tissue sample belonging to a DII patient, and the target phenotype as a matrix in which each row corresponded to the average cell type composition of tissues belonging to a CLR patient, and computed an optimal transport coupling between these two phenotypes (**Figure 4H**, **rows correspond to DII samples and columns to CLR samples**). We used this phenotypic edit, in combination with the above PSA with GRCh38, as an input to CPSA: a discrepancy was computed at each 128-mer by taking a sum (weighted by the computed coupling) of discrepancies between the pairs of functionals assigned to each pair of rows, one from the original and one from the target phenotype matrices, aggregated over genes as before (**see Section 3 of Supplementary Note 1 for an explanation of this procedure**).

The top 100 most discrepant genes were enriched for one GO biological process, "positive regulation of multicellular organismal process" (**Figure 4G, table**). Four of these genes were associated with B cell infiltrate in human tumors (**Figure 4G, underlined enriched genes**): the expression of FOXD1 was associated with B cell and DC activity in head and neck squamous carcinoma and associated with survival in CRC (Huang et al., 2021) and ADRB1 was associated with the B cell and DC infiltrate in breast cancer(Wang et al., 2020). Therapeutic inhibition of CXCR4 in CRC patients induced an integrated B cell response (Biasci et al., 2020). Thus, the discrepancy between the CLR and DII could be identifying sequences of causal genetic agents of the phenotypic edit, without any a priori biological definition of the of the precise CMP.

We further identified the genes that had a higher discrepancy in trained templates relative to untrained templates (**Supplementary Figure 3B, table**). Amongst the most discrepant genes (sixth overall), we found the zinc transporter SLC30A1, the high-expression of which was significantly associated with overall survival in Stage IV colorectal cancer patients

(**Supplementary Figure 3C, Kaplan-Meier Curve,** p = 0.01, TCGA data (Liu et al., 2018), courtesy of Human Protein Atlas (Uhlen et al., 2017)).

CPSA applied to two distinct edits to the CRC iTME revealed genomic loci corresponding to genetic modulators of the desired edits to the CMP, confirming that the identified PSAs of tissue organization with GRCh38 were, at least partially, templates. Moreover, these results show CPSA applied to PSAs obtained from phenotypic data alone and without biological annotations can lead to discovery of potential genetic therapeutic and diagnostic targets.

Discussion

In this work, we proposed and validated a theory describing how the information of a complex molecular phenotype (CMP) is specified by templates in the genome.

Our notion of template was a formal analogue of a coding-sequence of a protein and used a reformulation of the genome as a metric space of k-mer functionals. We algorithmically implemented a neural optimal transport algorithm, Neural Alignment of CMP with Reference genome (NACR), which aligns a measurement of a CMP with a genome sequence and outputs a phenotype sequence alignment (PSA). We further implemented causal phenotype sequence alignment (CPSA), which assigns a discrepancy functional to a CMP edit using a PSA. We defined a PSA to be a template when the discrepancy functional for any given edit to the CMP highlights the genomic loci whose manipulation would result in production of the edited CMP.

We found PSAs of CMPs from three fundamental biological contexts and showed that they were partially templates. Specifically, discrepancies for shifting gene expression in PBMCs as well as discrepancies for shifting lineage proportions in embryogenesis highlighted the sequences of genes that had been empirically established in perturbation studies to achieve the correct to the CMPs. Discrepancies for edits to the tissue organization of the CRC iTME further highlighted not only the sequences of genes but also of promoters and individual protein active sites whose empirical manipulation achieves the desired phenotypic changes.

The PSA of the CRC iTME with GRCh38 was weakly associated with colorectal SNPs. However, a limitation of this work is that we did not broadly evaluate how the genomic loci output by CPSA relate to loci learned from genome-wide association studies (van der Wijst et al., 2020; Yazar et al., 2022). Likely, the NACR algorithm could be enhanced by incorporation of population-scale measurements and biological annotations, akin to how prior knowledge of cellular localization was incorporated to reconstructing cellular spatial arrangements in (Nitzan et al., 2019) as well as by incorporation of other strategies for causal representation learning (Schölkopf et al., 2021). At the same time, conventional approaches for estimating causal effects that are applied to models for genotype-phenotype associations (Walker et al., 2022), or oracles for genotype-phenotype prediction (Vaishnav et al., 2022) may be enhanced by incorporating PSAs.

A further limitation of this work is that we did not address the theoretical or statistical properties of PSAs. Understanding the geometry and statistics of PSAs (existence, uniqueness and extending null alignments), could provide theoretically and statistically grounded strategies for identifying templates. Complementary directions for theoretical investigation include using the analogy described in **Supplementary Note 2** to search for templates of CMPs expressed in terms of other structures, such as formal languages (Bhate, 2021; Bhate et al., 2022); such studies may benefit from alternate formulations of systems (Döring and Isham, 2010). Furthermore, we adopted a weak definition of causality (requiring evaluation of only loci within the alignment). Alternative definitions of causality capturing how loci must be specifically changed, as well as comparison with traditional notions of genetic causality may provide enhanced understanding of CMPs and templates.

In NACR, we applied relatively simple neural network architectures. The recent success of self-attention based models over convolutional networks for modelling sequences (Le et al., 2021) raises the question of whether such architectures could also improve the ability to identify templates of CMPs.

Although we showed in simulations that information-efficiency constraints on the genome could result in CMPs having templates, a discussion of the biological and evolutionary mechanisms resulting in templates was not provided. If CMPs broadly have templates, could it provide a modular and direct strategy for organisms to efficiently and robustly adapt to selective pressures on those CMPs, and thus a potential mechanism for evolvability (Wagner and Zhang, 2011)? Recent work investigating the genotype-phenotype map in the context of algorithmic complexity theory (Johnston et al., 2022) and information acquisition approaches toward evolution (McGee et al., 2022) may suggest experimental strategies for such questions.

NACR requires only a single, rich phenotypic measurement to find PSAs. Our results therefore suggest a strategy for identifying genetic targets for diseases where small patient cohorts provide sample-size challenges for understandings the genetics of CMPs (Phillips et al., 2021). Beyond therapeutic targets, identifying genetic modulators of CMPs using CPSA may also provide mechanistic insights into disease biology.

Funding:

Funding to support this research was provided for by the Eric and Wendy Schmidt Center. AS was supported by the Society of Fellows at Harvard University. JC was supported by a Schmidt Fellowship at the Broad Institute of MIT and Harvard.

Author contributions:

SSB conceived the study, analyzed the theoretical framework, and conducted the computational experiments. AS analyzed the theoretical framework. JC supervised the study. All authors read and approved the manuscript.

Acknowledgements:

We would like to thank Dr. Orr Ashenberg (Broad Institute of MIT and Harvard), Professor Caroline Uhler (Broad Institute of MIT and Harvard and Massachusetts Institute of Technology), Professor Fei Chen (Broad Institute of MIT and Harvard and Harvard Medical School) and Dr.

Anthony Phillipakkis (Broad Institute of MIT and Harvard) for helpful suggestions during the course of this work.

Conflicting Interests

The authors declare that they have no conflicting interests with this work.

Data and Code availability:

All data was publicly available and accessed as described in the methods. All code is provided in the attached zip file and described in the “code description” section of the methods.

Methods

Glossary of terms

The following are informal; mathematical definitions are in **Supplementary Note 1**.

- **Template:** an object used by a system to specify the structure of a produced output; see below for definition of template for complex molecular phenotypes.
- **Complex molecular phenotype (CMP):** phenotypes represented as a collection of points in a (high-dimensional) metric space.
- **Phenotype space:** The ambient metric space in which the points of a CMP reside.
- **k-mer functional**: An assignment of a real number to each substring of length k in a genome.
- **Genome as a metric space (genome)**: the metric space whose points are k-mer functionals and whose distances are the squared differences between k-mer functionals, averaged across the positions of the genome as a sequence.
- **Phenotype-sequence alignment (PSA):** A function from the phenotype space of a CMP to a genome (as a metric space) that (approximately) preserves distances.
- **Neural alignment of CMP with reference genome (NACR algorithm):** The neural network optimal transport algorithm that finds an approximate alignment of a CMP with a reference genome by gradient descent.
- **Null alignment:** The PSA's obtained by randomly initializing but not training the neural networks of NACR.
- **Discrepancy functional for a phenotypic edit with respect to a PSA:** A k-mer functional whose value on a k-mer represents how important that k-mer is for the edit.
- **Gene/promoter/k-mer discrepancies (for a phenotypic edit, with respect to a PSA)**: the discrepancy functional (for the phenotypic edit, with respect to a PSA), averaged over genes/k-mers/promoter.
- **Causal phenotype sequence alignment (CPSA)**: The procedure whose input is an edit to a CMP and a PSA of the CMP with a genome and whose output is the discrepancy functional for the edit with respect to the PSA.

- **Template of CMP**: a PSA of a CMP with the reference genome of the organism producing the CMP whose causal predictions (as output by CPSA) have been experimentally confirmed.

Sequence data processing

Reference genome download

The utilized reference genomes were:

- GRCm38/mm10 for mouse (*M. musculus*) ("Genome Reference Consortium," n.d.)
- GRCh38/hg38 for human (Schneider et al., 2017)
- sacCer3 for yeast (*S. cerevisiae)*, downloaded from the Saccharomyces Genome Database (Cherry et al., 2012)
- NC_000913.3/U00096.3 for *E. coli* (Riley et al., 2006).

k-mer representations of genomes

FASTA files were downloaded from the aforementioned sources for each genome and manipulated with Biopython (Cock et al., 2009). Each downloaded chromosome was split into contigs that did not contain any 'N' nucleotides. Patches to the assembly were not included. Contigs were one-hot encoded, as a two dimensional array in which the index in the first dimension represented position in the sequence and the second dimension had a 1 in index 0, 1, 2 or 3 corresponding to A, T, C and G respectively. One-hot contigs were concatenated, keeping track of indices where 128-mers wholly contained within one contig could be extracted.

Except for the analysis of identifying models for the human tonsil in multiple species under differing regularization (detailed later in Methods), the human and mouse genome were represented by random extraction of 10,000,000 128-mers (each contained within one contig). Each trained model utilized the same random seed for

extraction, so any differences across datasets cannot be attributed to different starting genomes.

Gene sequences

Each gene symbol was represented by a collection of one-hot encoded 128-mers in two steps. The same procedure was utilized for the mouse and human genomes.

- First, a one-hot encoded transcript sequence gene sequence was extracted from the genome for each gene symbol. The RefSeq (O'Leary et al., 2016) table of genes was downloaded from the UCSC genome browser portal (Rosenbloom et al., 2015). Since multiple transcripts represented the same gene, an approximate transcription start and end site was obtained by averaging the start and end sites of the transcripts with the same gene symbol in the same chromosome. The sequence was one-hot encoded and internal N nucleotides were removed. Transcripts with coding sequences shorter than 1000 nucleotides were excluded in downstream analysis.
- The one-hot encoded sequence for each gene symbol was split into non-overlapping 128-mers starting at the zeroth position. This yielded a variable number of 128-mers for each gene.

Extraction of promoter, terminator, intergenic and transcript regions

In **Figure 2G**, functionals are evaluated on 128-mers extracted from different regions:

- The transcript regions were defined as regions within the txStart and txEnd positions of some gene in the RefSeq table.
- The promoter regions were defined as those within 2kb upstream of the txStart position of some gene in the RefSeq table on the + strand, or 2kb downstream of the txEnd position of some gene in the RefSeq table on the - strand.
- The terminator regions were defined as those within 2kb upstream of the txStart position of some gene in the RefSeq table on the - strand, or 2kb downstream of the txEnd position of some gene in the RefSeq table on the + strand.

- The intergenic regions were defined as those that were not promoter, transcript or terminator regions.

Promoter sequences

In **Figure 4**I, discrepancies are evaluated with respect to promoter sequences. Each gene symbol was represented by a collection of one-hot encoded 128-mers in two steps.

- First, a one-hot encoded promoter sequence was extracted from the genome for each gene symbol. Since multiple transcripts represented the same gene, promoters were extracted as the 2kb upstream of the RefSeq txStart position in the case of a + strand gene and 2kb downstream of the RefSeq txEnd position in the case of -strand gene. The sequence was one hot encoded and internal N nucleotides were removed.
- The one-hot encoded sequence for each gene symbol was split into non-overlapping 128-mers starting at the zeroth position. This yielded a variable number of 128-mers for each gene.
- When evaluating functionals, the aggregated promoter level values were averaged once more across gene symbols.

Phenotype data pre-processing

Each of the phenotypic datasets were the processed versions downloaded from prior publications and represented as a matrix of instances and features as detailed below.

Human tonsil CODEX dataset

The dataset was obtained from (Bhate et al., 2022; Kennedy-Darling et al., 2021). One tonsil from was utilized (referred to as tonsil 2 in (Bhate et al., 2022) and tonsil9338 in (Kennedy-Darling et al., 2021)). This dataset contained XY positions for each cell, cell type annotations and cellular neighborhood annotations. Each cell was represented by the total number of each cell type amongst its 30 nearest neighboring cells (computed using the scikit-learn package

(Pedregosa et al., 2011)) as previously described. The features were log transformed, adding a pseudocount of 0.01. A sample of 40000 cells was utilized in downstream analyses.

Colorectal Cancer (CRC) CODEX dataset

The dataset was obtained from (Schürch et al., 2020). This dataset contained XY positions for each cell, cell type, cellular neighborhood, tissue sample, patient and patient group annotations. Each cell was represented by the total number of each cell type amongst its 20 nearest neighboring cells (computed using the scikit-learn package) as previously described. The features were log transformed, adding a pseudocount of 1. A sample of 40000 cells was utilized in downstream analyses.

Peripheral Blood Mononuclear Cells (PBMC) single-cell gene expression dataset

The processed and filtered PBMC single-cell gene expression dataset was obtained from the Seurat data portal associated with (Hao et al., 2021). The h5seurat object was converted to the AnnData file format using the SeuratDisk package (Satija et al., 2019) and subsequently read using Scanpy (Wolf et al., 2018). Data were log-transformed, adding a pseudo-count of 1. Only the samples at the time 0 time-point (prior to any treatment) were utilized. The model used for **Figures 3F-H** was trained using the cells from the 0 time-point from patients 1, 3 and 5. The cell-type annotations used were the "celltype.l1" columns in all subsequent analysis.

Organogenesis single-cell gene expression dataset

The sample of 100000 processed and filtered single cells from the mouse organogenesis cell atlas (Cao et al., 2019) was obtained from the associated data portal. The data were log transformed, adding a pseudocount of 1. The 'lineage' annotations for each cell referred to in **Figure 3I** were the 'sub_trajectory_name' annotations.

NACR

Overview of approach

An overview of PSAs and the NACR algorithm is presented in **Supplementary Note 1, Sections 1-2.** The specific neural network architectures are detailed below. All models were built and trained using PyTorch (Paszke et al., 2019).

Methodology for analysis of models in genomes of different species

The one-hot encoded genome sequences were constructed according to the procedure described above. Next, 40 NACR models were trained at each regularization strength into each genome. In each replicate model, the initialized values of the neural network parameters were different.

- Genome: In each replicate model, a different sample of 10000 128-mers from the one-hot encoded genome for each species was used as the target genome. In the case of the random genome, 10000 128-mers were generated by uniform sampling.
- Phenotype: A sample of 40000 cells from the two tonsils in the HLT dataset was utilized, each represented by a 25 dimensional vector corresponding to the counts of each cell type within its 50 nearest neighbors.
- Latent dimension: A 4-dimensional latent inner-product space was utilized.
- Sequence embedding: Each 128-mer was represented as a 256-dimensional vector of counts of each possible 4-mer. The sequence-embedding component of phenotype-genotype embedding model was therefore a 256 x 4 affine map.
- Phenotype embedding: a 25x4 affine map was utilized.
- Regularization: L1 regularization with different strength was applied to the weights of the affine map (not the bias).
- Training: each model was trained for 10000 iterations with the Adam optimizer with learning rate 0.01. The sequence batch-size was 512, and the phenotype-instance batch-size was 256.

The alignment loss was estimated on a held-out evaluation dataset of 40000 cells not included in the training dataset as follows. It was not feasible to compute the deviation in the distances between pairs of cells and their corresponding functionals in the genome across all 40000x40000 possible pairs, and a batched estimate was therefore used.

The 40000 cells were split into evaluation batches of size 256. The distances between pairs of cells in each batch were computed in the original phenotype, as well as in the functional space, across the entire set of 10000 128-mers utilized for training (rescaling the functional outputs to adjust for different numbers of samples in the Monte-Carlo estimate across the whole genome). For each pair of cells in each evaluation batch, the maximum absolute discrepancy between their distances in the original phenotype space and in the genome were evaluated. Next, the maximum of these discrepancies were taken across batches as an evaluation of the overall model. This is a lower bound on the maximum across all possible pairs of cells, since a large number of, but not all, pairs are evaluated.

Architecture and training of NACR for human tonsil CODEX data

The following describes the NACR model used in **Figure 2E** to assess gene-level enrichment and Figure 2G to assess activations on 128-mers extracted from different regions:

- Genome: a sample of 10000000 128-mers from the one-hot encoded human genome was utilized.
- Phenotype: A sample of 40000 cells from the tonsil 2/tonsil9338 in the HLT dataset was utilized, each represented by a 25 dimensional vector corresponding to the log counts of each cell type within its 30 nearest neighbors (with a pseudocount of 0.01).
- Latent dimension: An 8-dimensional latent inner-product space was utilized. A ReLU activation was applied to the inner product to ensure positivity of the learned functionals.
- Sequence embedding. The sequence-embedding was a fully-convolutional network. The input was a (S, 4,128) tensor where S is the sequence batch-size. The layers of this network were:

    - Convolutional, input dim: 4, output dim: 16, kernel width: 5, padding: 2, stride 2, activation: ReLU, no bias
    - Max-pool, kernel-width: 2, stride: 2
    - Convolutional, input dim: 16, output dim: 8, kernel width: 5, padding: 2, stride 2, activation: ReLU, no bias
    - Max-pool, kernel-width: 2, stride: 2
    - Convolutional, input dim: 8, output dim: 8, kernel width: 5, padding: 2, stride 2, activation: ReLU, no bias
    - Max-pool, kernel-width: 2, stride: 2
    - Convolutional, input dim: 8, output dim: 8 (latent inner product space), kernel width:3, padding: 2, stride 2, activation: ReLU, no bias
    - Max-pool, kernel-width: 2, stride: 2.
    - Reshape to have output dimension(S,8)
- Phenotype embedding: a 3-layer perceptron architecture was used. The input was (B, 25) (where B was the phenotype-instance batch size, and 25 is the number of cell types in this dataset). The layers were:
    - Fully connected, input dim: 25, output dim 200, activation: ReLU.
    - Fully connected, input dim: 200, output dim 50 activation: ReLU.
    - Fully connected, input dim: 50, output dim 8 (latent inner product space) activation: ReLU.
- Regularization: A sparsity regularization of 0.00001 was applied to the absolute functional outputs summed across batches.
- Training: the model was trained for approximately 5 million iterations. The sequence batch-size was 512, and the phenotype-instance batch-size was 256.

Architecture and training of NACR for colorectal cancer CODEX data

The following describes the architecture and training of NACR on the CRC dataset used in **Figures 2F, 2H** and **Figure 4.**

- Genome: the same sample of 128-mers as the human tonsil.

- Phenotype: A sample of 40000 cells from the CRC dataset was utilized, each represented by a 29 dimensional vector corresponding to the log counts of each cell type within its 20 nearest neighbors (a pseudocount of 1 was added).
- Latent dimension: 8, with ReLU, same as HLT
- Sequence embedding: same as HLT.
- Phenotype embedding: same as HLT, except with input-dimension 29.
- Regularization: same as HLT.
- Training: same as HLT.

The same parameters and training procedure was utilized for each of the replicate models used to generate the results in **Figure 4B-I**.

Architecture and training of NACR on PBMC gene expression data

The following only describes the architecture and training of NACR on the PBMC dataset in **Figures 3F-H.** The methodology for assessment of the consistency of discrepancy scores across biological replicates is discussed after the general discussion of discrepancies.

- Genome: the same sample of 10000000 128-mers from the human genome as HLT/CRC dataset
- Phenotype: The log transformed expression values of cells at time 0 from donors 1, 3 and 5 in the dataset was utilized (a pseudocount of 1 was added).
- Latent dimension: 8, with ReLU, same as HLT/CRC.
- Sequence embedding: same as HLT/CRC.
- Phenotype embedding: same as HLT/CRC, except with input-dimension = 20729 human genes.
- Regularization: a sparsity regularization parameter of 0.00005 was utilized
- Training: same as HLT/CRC, except the model was trained for 2.5E7 iterations.

Architecture and training of NACR for mouse organogenesis single-cell gene expression data

The following only describes the architecture and training of NACR on the PBMC dataset in **Figures 3I.** The methodology for assessment of the computation of discrepancy scores is discussed after the general discussion of discrepancies.

- Genome: a sample of 10000000 128-mers from the mouse genome was utilized
- Phenotype: The log transformed expression values of cells was utilized (a pseudocount of 1 was added).
- Latent dimension: 8 with ReLU, same as HLT/CRC/PBMC.
- Sequence embedding: same as HLT/CRC/PBMC.
- Phenotype embedding: same as HLT/CRC/PBMC, except with input-dimension equal to the number of genes in the Cao et al dataset.
- Regularization: a sparsity regularization parameter of 0.000001, as well as a complexity regularization parameter (L2, 0.00001) were utilized.
- Training: same as HLT/CRC/PBMC, except the model was trained for 6E6 iterations.

Discrepancy analysis

Overview of discrepancy computations with NACR

A general overview of the discrepancy functional motivation and computation is presented in **Supplementary Note 1, Section 3**. The specific computations are detailed below.

Simulations to assess association between discrepancies and mutations

The goal of the simulation is to assess whether discrepancies, as described above, can highlight the genomic sequence positions whose changes achieve desired phenotypic changes.

The strategy has seven stages:

1. Choose an original phenotype and desired target phenotype.
    - The original phenotype is set to be an equal mixture of one dimensional Gaussians with centers at -1 and 1 respectively, represented by a sample of 100 points.
    - The target phenotype is set to be an equal mixture of one dimensional Gaussians with centers at -1 and 1.5 respectively (i.e., shifting the center at 1 by 0.5 in the positive direction)
2. Sample a random binary genome
    - The sampled genome has a length of 1000.
3. Identify a genotype-phenotype map Φ. This is a neural network, whose input is the sampled genome and output is original phenotype.
    - The neural network takes as input a 1000 binary vector (representing a binary genome), and outputs a 100 dimensional real vector (representing the phenotype)
    - The network is a multi-layer perceptron with two hidden layers of dimension 200 and 50 respectively, with no biases.
    - The network is trained to match the sampled genome to the sampled phenotype (a single sample), at differing regularizations on the weights at each layer. In general, the network can perfectly fit the input data, and this is desired.
4. Apply our NACR to find a PSA of the sampled phenotype with the sampled genome, T.
    - The genome is represented by enumeration of the frequencies of 3-mers within it.
    - The k-mer embedding is given by a linear map on k-mers
    - The latent dimension is 2
    - The phenotype embedding is a linear map from R to R^2
    - The genotype embedding is a linear map from R^3 to R^2
    - Utilize the learned genomic model T to compute discrepancies for each 3-mer in the sampled genome for the target phenotype.
    - The discrepancy for a 3-mer s is taken to be $(T(1)(s)-T(1.5)(s))^2$
5. Use Φ to compute the "true" distance to target phenotype obtained by 'flipping' the genome (from 1 to 0 or vice versa) at each position in the genome (one at a time).

- The new phenotype is taken as the output of Φ on the mutated genome.
- The distance to target-phenotype is computed as the 1-Wasserstein distance to a sample of 100 points from the target distribution (the mixture with shifted means). This is computed using the python optimal transport package (Flamary et al., n.d.).

6. Take a rolling average of the "true" distance to target phenotype over 3-mers and assess the (Spearman) correlation between the discrepancies.

There were 20 randomly sampled genomes utilized. For each sampled genome, regularization on Φ, and regularization on phenotype-genotype embedding, there were 20 phenotype-genotype embedding models trained. Each point in **Figure 2C** represents the median correlation across these 20 phenotype-genotype embedding models.

Biological reproducibility of discrepancies on PBMC gene expression data

We now describe the NACR parameters used to assess reproducibility of phenotype-genotype embedding across individual patients in the PBMC dataset:

- Genome: the same sample of 10000000 128-mers from the genome as described for HLT/CRC was utilized for all models. Each 128-mer was represented by the counts of each 4-mer, as for the NACR of tonsil with different genomes.
- Phenotype: principal component analysis (PCA) with 1000 components was performed on the log transformed (with pseudocount of 1) PBMC single-cell gene expression data using the scikit-learn implementation.
- Latent inner product dimension: 8. A ReLU activation was applied to the inner product to ensure positivity of the embeddings.
- Sequence-embedding: a linear layer (input dimension = 256 possible 4-mers, output dimension 8) is used.
- Phenotype embedding: a linear layer (input dimension = 1000 principal components, output dimension 8)

- Training: each NACR was trained for 50000 iterations using the Adam optimizer with a learning rate of 0.01. The sequence batch-size was 512 and the single-cell batch-size was 256.
- Regularization: PSAs were estimated at multiple combinations of sparsity and complexity regularization to investigate their effects on reproducibility.

50000 random 128-mers were extracted from transcript regions. After PSAs had been obtained, discrepancies at each of these 128-mers were computed for increasing cell type expression in the principal directions. Specifically, for each of $CD4^+$ T cells, Monocytes and $CD8^+$ T cells, the difference in values at each 128-mer between the functional assigned to the mean expression (in the first 1000 principal components) and the value of the functional assigned to the mean expression increased by half a standard deviation in each of the first two principal components. Spearman correlations were computed. In addition, null alignments were also evaluated for this discrepancy.

Computation of gene-level discrepancies for increasing IFNG expression in CD8 T cells

A PSA was trained as described earlier in the Methods on the PBMC gene expression data. The IFNG gene was detected in 4.3% $CD8^+$ T cells in the PBMC dataset with a mean expression of 0.025. We identified 223 genes (including IFNG) that were detected in between 4 and 5% of CD8+ T cells and with expression between 0.02 and 0.03, and computed discrepancies corresponding to multiplying the mean expression vector of CD8+ T cells by a factor of 2 (i.e., approximately squaring in counts space since models were computed in log space) in the directions corresponding to each of these genes individually. These discrepancies were aggregated over genes as described above, yielding a matrix of dimension (num_genes, 223) of discrepancies. The discrepancies for each target gene were z-normalized across genes (i.e., across rows). The dominant rank-1 signal in this matrix was removed by performing singular value decomposition (SVD) in pytorch, and reconstructing the matrix but with the first singular

value set to 0. The entries of the reconstructed matrix were referred to as the normalized discrepancies.

Association between discrepancies and CRISPR activation screening data

The CRISPR transcriptional activation screen results for IFNG expression in $CD8^+$ T cells was obtained from (Schmidt et al., 2022). The data there was filtered to only include the CRISPR activation screen for IFNG expression in CD8 T cells (other screens were included in the data table). Next, the common set of genes between those evaluated in the screen and those for which the discrepancies had been computed (a total of ~16000) had been computed were identified.

For each of the 223 target genes, for whose upregulation discrepancies had been computed, the difference in median normalized discrepancy between hits and non-hits were computed at different absolute z-score (available in the data downloaded from the original publication) thresholds for defining a hit and receiver-operator characteristic curves and AUC were computed using scikit-learn. Permutation tests were 1-sided tests conducted by permuting hit assignments 2000 times in the case of medians and 500 times in the case of AUCs.

The association of normalized IFNG upregulation discrepancies with gene hits was evaluated using multiple samples of size 20 from the set of 223 target upregulation genes for computation of normalized discrepancies to ensure it was not sensitive to the choice of genes we had used.

The Pearson correlations between all genes and IFNG were computed directly, and the AUC for prediction of hits using correlations were assessed. The normalized discrepancies for upregulating IFNG were adjusted for correlations taking residuals from a linear model across genes, whose exogenous variable was the normalized discrepancy for IFNG upregulation and whose endogenous variables were a constant and the correlation with IFNG. The model was estimated using the statsmodels python package. The adjusted normalized discrepancies were then evaluated as above for association with gene hits.

CD4 Effectorness Gene Upregulation discrepancy analysis

A list of CD4 T cell effectorness genes was obtained from(Cano-Gamez et al., 2020), and 205 present in the single-cell gene expression matrix of the PBMC dataset were identified. The discrepancy functionals corresponding to increasing the mean CD4+ T cell gene expression vector by log(1.5) in the direction of each effectorness gene individually were computed. The union of the top 100 most discrepant genes was taken for subsequent analysis. For each target effectorness gene, its discrepancies were z-normalized across genes.

Next, clustering was performed by the using the seaborn clustermap and scipy fcluster package with a maximum number of clusters set to 15. Each cluster that had more than five genes was evaluated for significantly enriched Gene Ontology Biological Process gene sets using the EnrichR api via the GSEApy package (Kuleshov et al., 2016). Select gene sets with FDR < 0.1 (in each cluster individually, not overall) were visualized alongside the heatmap.

Lineage proportion increase discrepancy analysis

Modules of discrepant genes

A PSAl obtained as described above for the mouse organogenesis dataset with the mouse genome. The mean gene expression vectors and overall frequencies were computed for each lineage (corresponding to the 'sub_trajectory_name' annotation in the original dataset). For each lineage, a coupling matrix between the original frequencies and the frequencies corresponding to that lineage increased by a factor of 1.5 (with other marginals remaining constant) was computed. Specifically, entropy-regularized optimal transport using the Sinkhorn algorithm with a regularization parameter of 0.08 using the pairwise distance between cluster means as the cost matrix was computed using the python optimal transport software package. Next, for each 128-mer in the gene sequence data, a matrix of squared pairwise differences between the functionals assigned by the genomic model to each lineage gene expression mean vector were computed. An overall discrepancy for the 128-mer was computed by taking the

inner product between the matrix of squared pairwise functional differences and the optimal transport coupling. These discrepancies were aggregated over genes as described above.

The union of the 200 genes that were most discrepant for each target lineage increase were selected for downstream analysis, and discrepancies for each target lineage increase were z-normalized across these 200 genes. Next, clustering and evaluation of significantly enriched mouse GO BP categories was conducted as with the PBMC dataset.

Lineage-specific discrepant genes by regression

We reported above certain genes enriched with high discrepancy specific to inhibitory neurons, Excitatory neurons and White blood cell lineages. These were obtained by taking the vector of z-normalized gene discrepancies and estimating a linear model with covariates given by the discrepancies for the other lineages (estimated using the statsmodels python package), and subsequently ranking genes by their residuals with respect to this model. EnrichR was used to identify developmental genes within the top 10 ranked by these residuals.

Discrepancies for increasing intratumoral CD8 T cells in CRC

Gene level analysis

The average cell type frequencies in CN 2 (bulk tumor) were computed using annotations from (Schürch et al., 2020). Next, the discrepancies for increasing the counts of CD8+ T cells from 0.12 to 0.5 in log space were computed and aggregated across genes using 6 PSAs (training NACR as described above) and 99 null alignments (with the same architecture as the trained models). The proportion of pairs of trained models with Spearman correlation > 0.3 across genes was compared to pairs of untrained models via permutation of the label assignments.

Student's t test statistics and p-values were computed for the difference between the discrepancies at each gene between the 6 trained models and the 99 untrained models. These p-values were adjusted using the Benjamini-Hochberg procedure implemented in the

statsmodel package suggesting CXCL11 as the only significantly different one. The top twenty genes with highest test statistic are used as "hits" for the target phenotype and visualized in **Figure 4C**.

The TCGA-COADREAD gene expression dataset was downloaded from (Liu et al., 2018). The Spearman correlation and p-values of each of the hit genes with CD8A expression in this dataset was computed using the Scipy implementation. Significance was determined as correlation p value < 0.05/20 (i.e., Bonferroni corrected).

A null distribution was obtained to assess the significance of 15/20 being correlated with CD8A expression. Specifically, 6 of the 99 null alignments were randomly selected, and the t-test statistics were computed comparing the averaged discrepancies in those 6 null alignments to the 93 other null alignments to yield 20 null hit genes. The proportion of these that were significantly associated with CD8A expression was stored, and the procedure was repeated 5000 times. The reported p-value is the fraction of times that greater than or equal to 15/20 hit genes were significantly associated with CD8A expression. In addition, a binomial p value was computed using n = 20, p = 0.53 (which was the proportion of genes significantly correlated with CD8A).

Analysis along CXCL11 transcript

Each 128-mer whose 65^{th} base pair (centered at) was within the transcription start and end site of the CXCL11 transcript was extracted (GRCh38 assembly, using RefSeq annotations). Next, the discrepancies for increasing CD8+ T cell frequency in the tumor CN as above were computed for each of the six PSAs and 100 null alignments at each 128-mer were computed and z-normalized across all the 128-mers. A p-value for an increase in discrepancy at each position was obtained as follows. The difference between means at each 128-mer centered at each position in the transcript were computed, and a one-sided permutation p-value was computed by applying 20000 permutations to the untrained vs trained label. Next, the coding region within the third exon was translated using the Biopython translate function.

The score assigned to each amino acid was the minimum p-value assigned to the three positions contained within the codon coding for it.

Analysis of individual PSA discrepancies for CD8$^+$ T cells

The individual PSA used for **Figures 2F-I** was utilized. The top 100 most discrepant genes for the target (increase intratumoral CD8+ T cells from 0.14 to 0.5) were selected, and EnrichR analysis was performed for identification of significantly enriched GOBP categories (FDR < 0.1). Only two categories (those displayed in **Figure 4E**) were significantly enriched. The TCGA COAD-READ dataset was used via the cBioPortal to produce Kaplan-Meier plots for survival analysis and computation of log-rank p-values for alterations in the CD200R1, PTGS2 or MMP3. The other enriched category (with genes ROMO1, DEFA5, CFHR1) was assessed by manual literature search.

Analysis of promoter-level discrepancies

For each transcript in the RefSeq table, discrepancies were computed by extracting the 128-mers contained within the 5000bp upstream of the transcription start site (adjusting for the strand). Next, discrepancy functionals for increasing CD8+ T cell frequencies in CN2 were computed for the individual PSA used for **Figures 2F-I** and **Figures 4E-H** as above and aggregated over the promoter regions. Next, the promoter-level discrepancies were averaged across gene-symbols and genes were ranked by discrepancies for their promoters. EnrichR analysis was performed on the top 100 most discrepant genes, and only one GO BP category was significantly enriched (FDR < 0.1) as depicted in **Figure 4I**.

Discrepancies for architectural alteration

Tissue samples that did not contain CN5 (the follicle, defining characteristic of the patient groups) were selected in each patient group (51 for CLR, 67 for DII). The mean cell type frequency vectors for each sample was computed. The pairwise distance between these vectors

was used as the cost matrix to identify an optimal transport coupling between the distributions with equal weights supported at each of these vectors, using the implementation of the Sinkhorn algorithm (with regularization parameter = 0.1) in the python OT software package. Next, the pairwise differences between the functionals assigned by the PSAI used for **Figures 4E-I** to each vector were computed at each 128-mer and the inner product between this matrix and the learned optimal transport coupling were taken to yield a discrepancy for each 128-mer. These discrepancies were aggregated across genes. The top 100 most discrepant genes were selected, and enriched GOBP categories were identified using the EnrichR api implemented in the GSEApy package. The only enriched term is depicted in **Figure 4K**. The genes in this category were investigated by manual literature search.

Other biological analyses

Gene Set Enrichment Analysis

For Figures 2E and Figure 2F, gene set enrichment analysis was performed. First, functionals were computed for each 128-mer from the PSAs specified above:

- For the tonsil dataset, the functional used to rank genes was the difference in mean cell type frequency vectors across cells assigned to the dark zone and across cells assigned to the light zone. The CN assignments were obtained from (Bhate et al., 2022).
- For the CRC dataset, the functional used was the difference in mean cell type frequency vectors across patches assigned to the CLR patient group and across patches assigned to the DII patient group.

Next, a gene-level functional value was computed by averaging across 128-mers in the gene dataset, as described in the sequence data preparation section of Methods. Genes were ranked by functional values aggregated over genes, and enriched Gene Ontology Biological Process categories were identified using the fgsea R package (Korotkevich et al., 2021) (considering categories with between 10 and 500 genes). Those categories with FDR < 0.1 were selected.

A p-value was computed as the frequency of times each category was significantly enriched in 99 untrained models; the 'Train' p-value reported in **Figures 2E-F** are the Benjamini-Hochberg adjusted values.

Null alignments were used to obtain a p-value corresponding to how frequently each category that was significantly was enriched according to a functional obtained from PSA was enriched in a functional obtained from a null alignment and FDR adjusted.

SNP analysis

The dataset of pathogenic SNPs were downloaded from the ClinVar database (Landrum et al., 2018), as follows. First, a search for variants containing the term 'colorectal' was conducted. Variants labelled "pathogenic", "single-nucleotide" and "expert panel" validated were selected and downloaded. The difference in the functional assigned to PSA to the CLR cell type frequency mean and the DII cell type frequency mean was evaluated at the 128-mers centered at each SNP. SNPs with 'N' nucleotides in the GRCh38 assembly in their 128-mers were excluded, yielding a total of 419 variants evaluated in the subsequent analysis.

In addition to the 128-mers centered at each SNP, 1000 128-mers sampled at random from the 1MB either side of each SNP were extracted to form a dataset of 'null' variants. A two-sided permutation test with 1000 permutations was conducted to assess the presence of a difference in activations at the SNPs vs the null variants. Next, the same p-value was computed using 100 null alignments.

Description of Code

The following table describes the notebooks included with this publication. These contain the code and resulting figures as described.

Submission notebook name	Figures	Description
Notebook_1_sequence-data-preparation.ipynb	All	Generating 1hot encoded sequence data for genomes and genes
Notebook_2_Species-Comparison.ipynb	2C	Comparing PSA of tonsil with different genomes
Notebook_3_PG-embedding-HLT.ipynb	2E	NACR of tonsil with GRCh38
Notebook_4_PG-embedding-CRC.ipynb	2F	NACR of CRC iTME with GRCh38
Notebook_5-PG-embedding-CRC-multiple-initializations.ipynb	4B-D	NACR of CRC iTME with GRCh38 from multiple initializations
Notebook_6-PG-embedding-HLT-multiple-initializations.ipynb	not shown	NACR of tonsil with GRCh38 from multiple initializations
Notebook_7_CRC-model-analysis-part1.ipynb	2F, 4B-D	Gene-level scores for CLR/DII difference functional with PSA and null alignments; gene and AA level discrepancies for $CD8^+$ T cell increases
Notebook_8-HLT-model-analysis.ipynb	2E	Gene-level scores for Dark Zone/Light Zone difference functional with PSA and null alignments
Notebook_9_HLT-GSEA-R.ipynb	2E	GSEA analysis in R for DarkZone/Light Zone difference
Notebook_10_CRC-GSEA-R.ipynb	2F	GSEA analysis in R for CLR/DII difference functional
Notebook_11_comparing-region-activations.ipynb	2G	Comparing functionals values evaluated on different types of regions
Notebook_12_CRC-model-evaluation-part2.ipynb	2H, 4E-K	Comparing individual CRC DII vs CLR on pathogenic SNPs, intratumoral $CD8^+$ and CLR vs DII discrepancies
Notebook_13_simulation.ipynb	3C	Simulating genotype-phenotype map to measure correlations with discrepancies and distance to target phenotype
Notebook_14_PBMC-biological-replicates.ipynb	3E	Biological replicates correlations for 1-layer NACR between discrepancies for increasing expression of PCs in cell types

Notebook_15_PG-embedding-PBMC.ipynb	3F-H	NACR of PBMC dataset with GRCh38
Notebook_16_PBMC-model-analysis.ipynb	3F-H	Analysis of discrepancies for IFNG and CD4 effectorness
Notebook_17_MOCA-biological-replicates.ipynb	not used	Biological replicates correlations for 1-layer models between discrepancies for increasing expression of PCs in cell types
Notebook_18_PG-embedding-MOCA.ipynb	3I	NACR of mammalian organogenesis dataset into mm10
Notebook_19_MOCA-model-analysis.ipynb	3I	Analysis of discrepancies for increasing embryo lineages

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. https://doi.org/10.1038/75556

Bhate, S.S., 2021. Towards semantic representations of tissue organization from high-parameter imaging data. Stanford University.

Bhate, S.S., Barlow, G.L., Schürch, C.M., Nolan, G.P., 2022. Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors. Cell Syst. 13, 109-130.e6. https://doi.org/10.1016/j.cels.2021.09.012

Biasci, D., Smoragiewicz, M., Connell, C.M., Wang, Z., Gao, Y., Thaventhiran, J.E.D., Basu, B., Magiera, L., Johnson, T.I., Bax, L., Gopinathan, A., Isherwood, C., Gallagher, F.A., Pawula, M., Hudecova, I., Gale, D., Rosenfeld, N., Barmpounakis, P., Popa, E.C., Brais, R., Godfrey, E., Mir, F., Richards, F.M., Fearon, D.T., Janowitz, T., Jodrell, D.I., 2020. CXCR4 inhibition in human pancreatic and colorectal cancers induces an integrated immune response. Proc. Natl. Acad. Sci. 117, 28960–28970. https://doi.org/10.1073/pnas.2013644117

Bruni, D., Angell, H.K., Galon, J., 2020. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. Nat. Rev. Cancer 20, 662–680. https://doi.org/10.1038/s41568-020-0285-7

Cano-Gamez, E., Soskic, B., Roumeliotis, T.I., So, E., Smyth, D.J., Baldrighi, M., Willé, D., Nakic, N., Esparza-Gordillo, J., Larminie, C.G.C., Bronson, P.G., Tough, D.F., Rowan, W.C., Choudhary, J.S., Trynka, G., 2020. Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines. Nat. Commun. 11, 1801. https://doi.org/10.1038/s41467-020-15543-y

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., Trapnell, C., Shendure, J., 2019. The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496–502. https://doi.org/10.1038/s41586-019-0969-x

Cao, Y., Jiao, N., Sun, T., Ma, Y., Zhang, X., Chen, H., Hong, J., Zhang, Y., 2021. CXCL11 Correlates With Antitumor Immunity and an Improved Prognosis in Colon Cancer. Front. Cell Dev. Biol. 9.

Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Karra, K., Krieger, C.J., Miyasato, S.R., Nash, R.S., Park, J., Skrzypek, M.S., Simison, M., Weng, S., Wong, E.D., 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 40, D700–D705. https://doi.org/10.1093/nar/gkr1029

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., Zhang, F., 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science 339, 819–823. https://doi.org/10.1126/science.1231143

Crick, F., 1970. Central Dogma of Molecular Biology. Nature 227, 561–563. https://doi.org/10.1038/227561a0

Davidson, D.J., Currie, A.J., Reid, G.S.D., Bowdish, D.M.E., MacDonald, K.L., Ma, R.C., Hancock, R.E.W., Speert, D.P., 2004. The Cationic Antimicrobial Peptide LL-37 Modulates Dendritic Cell Differentiation and Dendritic Cell-Induced T Cell Polarization. J. Immunol. 172, 1146–1156. https://doi.org/10.4049/jimmunol.172.2.1146

Döring, A., Isham, C., 2010. "What is a thing?": topos theory in the foundations of physics, in: New Structures for Physics. Springer, pp. 753–937.

Fessler, E., Medema, J.P., 2016. Colorectal Cancer Subtypes: Developmental Origin and Microenvironmental Regulation. Trends Cancer 2, 505–518. https://doi.org/10.1016/j.trecan.2016.07.008

Flamary, R., Courty, N., Gramfort, A., Alaya, M.Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N.T.H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D.J., Tavenard, R., Tong, A., Vayer, T., n.d. POT: Python Optimal Transport 8.

Flore, G., Cioffi, S., Bilio, M., Illingworth, E., 2017. Cortical Development Requires Mesodermal Expression of Tbx1, a Gene Haploinsufficient in 22q11.2 Deletion Syndrome. Cereb. Cortex 27, 2210–2225. https://doi.org/10.1093/cercor/bhw076

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., Schultz, N., 2013. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci. Signal. 6, pl1–pl1. https://doi.org/10.1126/scisignal.2004088

Genome Reference Consortium [WWW Document], n.d. URL https://www.ncbi.nlm.nih.gov/grc (accessed 4.11.22).

Grädel, E., Kolaitis, P.G., Libkin, L., Marx, M., Spencer, J., Vardi, M.Y., Venema, Y., Weinstein, S., 2007. Finite Model Theory and its applications. Springer Science & Business Media.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R., 2021. Integrated analysis of multimodal single-cell data. Cell 184, 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

Hodis, E., Triglia, E.T., Kwon, J.Y.H., Biancalani, T., Zakka, L.R., Parkar, S., Hütter, J.-C., Buffoni, L., Delorey, T.M., Phillips, D., Dionne, D., Nguyen, L.T., Schapiro, D., Maliga, Z., Jacobson, C.A., Hendel, A., Rozenblatt-Rosen, O., Mihm, M.C., Garraway, L.A., Regev, A., 2022. Stepwise-edited, human melanoma models reveal mutations' effect on tumor and microenvironment. Science 376, eabi8175. https://doi.org/10.1126/science.abi8175

Hou, P.-S., hAilín, D.Ó., Vogel, T., Hanashima, C., 2020. Transcription and Beyond: Delineating FOXG1 Function in Cortical Development and Disorders. Front. Cell. Neurosci. 14.

Huang, J., Liang, B., Wang, T., 2021. FOXD1 expression in head and neck squamous carcinoma: a study based on TCGA, GEO and meta-analysis. Biosci. Rep. 41, BSR20210158. https://doi.org/10.1042/BSR20210158

Huang, W., Li, L., Myers, J.R., Marth, G.T., 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28, 593–594. https://doi.org/10.1093/bioinformatics/btr708

Ise, W., Kurosaki, T., 2019. Plasma cell differentiation during the germinal center reaction. Immunol. Rev. 288, 64–74. https://doi.org/10.1111/imr.12751

Johnston, I.G., Dingle, K., Greenbury, S.F., Camargo, C.Q., Doye, J.P.K., Ahnert, S.E., Louis, A.A., 2022. Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution. Proc. Natl. Acad. Sci. 119, e2113883119. https://doi.org/10.1073/pnas.2113883119

Kennedy-Darling, J., Bhate, S.S., Hickey, J.W., Black, S., Barlow, G.L., Vazquez, G., Venkataraaman, V.G., Samusik, N., Goltsev, Y., Schürch, C.M., others, 2021. Highly multiplexed tissue imaging using repeated oligonucleotide exchange reaction. Eur. J. Immunol.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., Teichmann, S.A., 2015. The technology and biology of single-cell RNA sequencing. Mol. Cell 58, 610–620.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., Sergushichev, A., 2021. Fast gene set enrichment analysis. https://doi.org/10.1101/060012

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., McDermott, M.G., Monteiro, C.D., Gundersen, G.W., Ma'ayan, A., 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 44, W90–W97. https://doi.org/10.1093/nar/gkw377

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J.B., Kattman, B.L., Maglott, D.R., 2018. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 46, D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Le, N.Q.K., Ho, Q.-T., Nguyen, T.-T.-D., Ou, Y.-Y., 2021. A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. Brief. Bioinform. 22, bbab005. https://doi.org/10.1093/bib/bbab005

Liang, M., Wang, J., Wu, C., Wu, M., Hu, J., Dai, J., Ruan, H., Xiong, S., Dong, C., 2021. Targeting matrix metalloproteinase MMP3 greatly enhances oncolytic virus mediated tumor therapy. Transl. Oncol. 14, 101221. https://doi.org/10.1016/j.tranon.2021.101221

Liu, Jianfang, Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., Omberg, L., Wolf, D.M., Shriver, C.D., Thorsson, V., Hu, H., Caesar-Johnson, S.J., Demchok, J.A., Felau, I., Kasapi, M., Ferguson, M.L., Hutter, C.M., Sofia, H.J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J.C., Zhang, J. (Julia), Chudamani, S., Liu, Jia, Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Cho, J., DeFreitas, T., Frazer, S., Gehlenborg, N., Getz, G., Heiman, D.I., Kim, J., Lawrence, M.S., Lin, P., Meier, S., Noble, M.S., Saksena, G., Voet, D., Zhang, Hailei, Bernard, B., Chambwe, N., Dhankani, V., Knijnenburg, T., Kramer, R., Leinonen, K., Liu, Y., Miller, M., Reynolds, S., Shmulevich, I., Thorsson, V., Zhang, W., Akbani, R., Broom, B.M., Hegde, A.M., Ju, Z., Kanchi, R.S., Korkut, A., Li, J., Liang, H., Ling, S., Liu, W., Lu, Y., Mills, G.B., Ng, K.-S., Rao, A., Ryan, M., Wang, Jing, Weinstein, J.N., Zhang, J., Abeshouse, A., Armenia, J., Chakravarty, D., Chatila, W.K., de Bruijn, I., Gao, J., Gross, B.E., Heins, Z.J., Kundra, R.,

La, K., Ladanyi, M., Luna, A., Nissan, M.G., Ochoa, A., Phillips, S.M., Reznik, E., Sanchez-Vega, F., Sander, C., Schultz, N., Sheridan, R., Sumer, S.O., Sun, Y., Taylor, B.S., Wang, Jioajiao, Zhang, Hongxin, Anur, P., Peto, M., Spellman, P., Benz, C., Stuart, J.M., Wong, C.K., Yau, C., Hayes, D.N., Parker, J.S., Wilkerson, M.D., Ally, A., Balasundaram, M., Bowlby, R., Brooks, D., Carlsen, R., Chuah, E., Dhalla, N., Holt, R., Jones, S.J.M., Kasaian, K., Lee, D., Ma, Y., Marra, M.A., Mayo, M., Moore, R.A., Mungall, A.J., Mungall, K., Robertson, A.G., Sadeghi, S., Schein, J.E., Sipahimalani, P., Tam, A., Thiessen, N., Tse, K., Wong, T., Berger, A.C., Beroukhim, R., Cherniack, A.D., Cibulskis, C., Gabriel, S.B., Gao, G.F., Ha, G., Meyerson, M., Schumacher, S.E., Shih, J., Kucherlapati, M.H., Kucherlapati, R.S., Baylin, S., Cope, L., Danilova, L., Bootwalla, M.S., Lai, P.H., Maglinte, D.T., Van Den Berg, D.J., Weisenberger, D.J., Auman, J.T., Balu, S., Bodenheimer, T., Fan, C., Hoadley, K.A., Hoyle, A.P., Jefferys, S.R., Jones, C.D., Meng, S., Mieczkowski, P.A., Mose, L.E., Perou, A.H., Perou, C.M., Roach, J., Shi, Y., Simons, J.V., Skelly, T., Soloway, M.G., Tan, D., Veluvolu, U., Fan, H., Hinoue, T., Laird, P.W., Shen, H., Zhou, W., Bellair, M., Chang, K., Covington, K., Creighton, C.J., Dinh, H., Doddapaneni, H., Donehower, L.A., Drummond, J., Gibbs, R.A., Glenn, R., Hale, W., Han, Y., Hu, J., Korchina, V., Lee, S., Lewis, L., Li, W., Liu, X., Morgan, M., Morton, D., Muzny, D., Santibanez, J., Sheth, M., Shinbro, E., Wang, L., Wang, M., Wheeler, D.A., Xi, L., Zhao, F., Hess, J., Appelbaum, E.L., Bailey, M., Cordes, M.G., Ding, L., Fronick, C.C., Fulton, L.A., Fulton, R.S., Kandoth, C., Mardis, E.R., McLellan, M.D., Miller, C.A., Schmidt, H.K., Wilson, R.K., Crain, D., Curley, E., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Sherman, M., Thompson, E., Yena, P., Bowen, J., Gastier-Foster, J.M., Gerken, M., Leraas, K.M., Lichtenberg, T.M., Ramirez, N.C., Wise, L., Zmuda, E., Corcoran, N., Costello, T., Hovens, C., Carvalho, A.L., de Carvalho, A.C., Fregnani, J.H., Longatto-Filho, A., Reis, R.M., Scapulatempo-Neto, C., Silveira, H.C.S., Vidal, D.O., Burnette, A., Eschbacher, J., Hermes, B., Noss, A., Singh, R., Anderson, M.L., Castro, P.D., Ittmann, M., Huntsman, D., Kohl, B., Le, X., Thorp, R., Andry, C., Duffy, E.R., Lyadov, V., Paklina, O., Setdikova, G., Shabunin, A., Tavobilov, M., McPherson, C., Warnick, R., Berkowitz, R., Cramer, D., Feltmate, C., Horowitz, N., Kibel, A., Muto, M., Raut, C.P., Malykh, A., Barnholtz-Sloan, J.S., Barrett, W., Devine, K., Fulop, J., Ostrom, Q.T., Shimmel, K., Wolinsky, Y., Sloan, A.E., De Rose, A., Giuliante, F., Goodman, M., Karlan, B.Y., Hagedorn, C.H., Eckman, J., Harr, J., Myers, J., Tucker, K., Zach, L.A., Deyarmin, B., Hu, H., Kvecher, L., Larson, C., Mural, R.J., Somiari, S., Vicha, A., Zelinka, T., Bennett, J., Iacocca, M., Rabeno, B., Swanson, P., Latour, M., Lacombe, L., Têtu, B., Bergeron, A., McGraw, M., Staugaitis, S.M., Chabot, J., Hibshoosh, H., Sepulveda, A., Su, T., Wang, T., Potapova, O., Voronina, O., Desjardins, L., Mariani, O., Roman-Roman, S., Sastre, X., Stern, M.-H., Cheng, F., Signoretti, S., Berchuck, A., Bigner, D., Lipp, E., Marks, J., McCall, S., McLendon, R., Secord, A., Sharp, A., Behera, M., Brat, D.J., Chen, A., Delman, K., Force, S., Khuri, F., Magliocca, K., Maithel, S., Olson, J.J., Owonikoko, T., Pickens, A., Ramalingam, S., Shin, D.M., Sica, G., Van Meir, E.G., Zhang, Hongzheng, Eijckenboom, W., Gillis, A., Korpershoek, E., Looijenga, L., Oosterhuis, W., Stoop, H., van Kessel, K.E., Zwarthoff, E.C., Calatozzolo, C., Cuppini, L., Cuzzubbo, S., DiMeco, F., Finocchiaro, G., Mattei, L., Perin, A., Pollo, B., Chen, C., Houck, J., Lohavanichbutr, P., Hartmann, A., Stoehr, C., Stoehr, R., Taubert, H., Wach, S., Wullich, B., Kycler, W., Murawa, D., Wiznerowicz, M., Chung, K., Edenfield, W.J., Martin, J.,

Baudin, E., Bubley, G., Bueno, R., De Rienzo, A., Richards, W.G., Kalkanis, S., Mikkelsen, T., Noushmehr, H., Scarpace, L., Girard, N., Aymerich, M., Campo, E., Giné, E., Guillermo, A.L., Van Bang, N., Hanh, P.T., Phu, B.D., Tang, Y., Colman, H., Evason, K., Dottino, P.R., Martignetti, J.A., Gabra, H., Juhl, H., Akeredolu, T., Stepa, S., Hoon, D., Ahn, K., Kang, K.J., Beuschlein, F., Breggia, A., Birrer, M., Bell, D., Borad, M., Bryce, A.H., Castle, E., Chandan, V., Cheville, J., Copland, J.A., Farnell, M., Flotte, T., Giama, N., Ho, T., Kendrick, M., Kocher, J.-P., Kopp, K., Moser, C., Nagorney, D., O'Brien, D., O'Neill, B.P., Patel, T., Petersen, G., Que, F., Rivera, M., Roberts, L., Smallridge, R., Smyrk, T., Stanton, M., Thompson, R.H., Torbenson, M., Yang, J.D., Zhang, L., Brimo, F., Ajani, J.A., Angulo Gonzalez, A.M., Behrens, C., Bondaruk, J., Broaddus, R., Czerniak, B., Esmaeli, B., Fujimoto, J., Gershenwald, J., Guo, C., Lazar, A.J., Logothetis, C., Meric-Bernstam, F., Moran, C., Ramondetta, L., Rice, D., Sood, A., Tamboli, P., Thompson, T., Troncoso, P., Tsao, A., Wistuba, I., Carter, C., Haydu, L., Hersey, P., Jakrot, V., Kakavand, H., Kefford, R., Lee, K., Long, G., Mann, G., Quinn, M., Saw, R., Scolyer, R., Shannon, K., Spillane, A., Stretch, J., Synott, M., Thompson, J., Wilmott, J., Al-Ahmadie, H., Chan, T.A., Ghossein, R., Gopalan, A., Levine, D.A., Reuter, V., Singer, S., Singh, B., Tien, N.V., Broudy, T., Mirsaidi, C., Nair, P., Drwiega, P., Miller, J., Smith, J., Zaren, H., Park, J.-W., Hung, N.P., Kebebew, E., Linehan, W.M., Metwalli, A.R., Pacak, K., Pinto, P.A., Schiffman, M., Schmidt, L.S., Vocke, C.D., Wentzensen, N., Worrell, R., Yang, H., Moncrieff, M., Goparaju, C., Melamed, J., Pass, H., Botnariuc, N., Caraman, I., Cernat, M., Chemencedji, I., Clipca, A., Doruc, S., Gorincioi, G., Mura, S., Pirtac, M., Stancul, I., Tcaciuc, D., Albert, M., Alexopoulou, I., Arnaout, A., Bartlett, J., Engel, J., Gilbert, S., Parfitt, J., Sekhon, H., Thomas, G., Rassl, D.M., Rintoul, R.C., Bifulco, C., Tamakawa, R., Urba, W., Hayward, N., Timmers, H., Antenucci, A., Facciolo, F., Grazi, G., Marino, M., Merola, R., de Krijger, R., Gimenez-Roqueplo, A.-P., Piché, A., Chevalier, S., McKercher, G., Birsoy, K., Barnett, G., Brewer, C., Farver, C., Naska, T., Pennell, N.A., Raymond, D., Schilero, C., Smolenski, K., Williams, F., Morrison, C., Borgia, J.A., Liptay, M.J., Pool, M., Seder, C.W., Junker, K., Omberg, L., Dinkin, M., Manikhas, G., Alvaro, D., Bragazzi, M.C., Cardinale, V., Carpino, G., Gaudio, E., Chesla, D., Cottingham, S., Dubina, M., Moiseenko, F., Dhanasekaran, R., Becker, K.-F., Janssen, K.-P., Slotta-Huspenina, J., Abdel-Rahman, M.H., Aziz, D., Bell, S., Cebulla, C.M., Davis, A., Duell, R., Elder, J.B., Hilty, J., Kumar, B., Lang, J., Lehman, N.L., Mandt, R., Nguyen, P., Pilarski, R., Rai, K., Schoenfield, L., Senecal, K., Wakely, P., Hansen, P., Lechan, R., Powers, J., Tischler, A., Grizzle, W.E., Sexton, K.C., Kastl, A., Henderson, J., Porten, S., Waldmann, J., Fassnacht, M., Asa, S.L., Schadendorf, D., Couce, M., Graefen, M., Huland, H., Sauter, G., Schlomm, T., Simon, R., Tennstedt, P., Olabode, O., Nelson, M., Bathe, O., Carroll, P.R., Chan, J.M., Disaia, P., Glenn, P., Kelley, R.K., Landen, C.N., Phillips, J., Prados, M., Simko, J., Smith-McCune, K., VandenBerg, S., Roggin, K., Fehrenbach, A., Kendler, A., Sifri, S., Steele, R., Jimeno, A., Carey, F., Forgie, I., Mannelli, M., Carney, M., Hernandez, B., Campos, B., Herold-Mende, C., Jungk, C., Unterberg, A., von Deimling, A., Bossler, A., Galbraith, J., Jacobus, L., Knudson, M., Knutson, T., Ma, D., Milhem, M., Sigmund, R., Godwin, A.K., Madan, R., Rosenthal, H.G., Adebamowo, C., Adebamowo, S.N., Boussioutas, A., Beer, D., Giordano, T., Mes-Masson, A.-M., Saad, F., Bocklage, T., Landrum, L., Mannel, R., Moore, K., Moxley, K., Postier, R., Walker, J., Zuna, R., Feldman, M., Valdivieso, F., Dhir, R., Luketich, J., Mora Pinero, E.M.,

Quintero-Aguilo, M., Carlotti, Jr, C.G., Dos Santos, J.S., Kemp, R., Sankarankuty, A., Tirapelli, D., Catto, J., Agnew, K., Swisher, E., Creaney, J., Robinson, B., Shelley, C.S., Godwin, E.M., Kendall, S., Shipman, C., Bradford, C., Carey, T., Haddad, A., Moyer, J., Peterson, L., Prince, M., Rozek, L., Wolf, G., Bowman, R., Fong, K.M., Yang, I., Korst, R., Rathmell, W.K., Fantacone-Campbell, J.L., Hooke, J.A., Kovatich, A.J., Shriver, C.D., DiPersio, J., Drake, B., Govindan, R., Heath, S., Ley, T., Van Tine, B., Westervelt, P., Rubin, M.A., Lee, J.I., Aredes, N.D., Mariamidze, A., 2018. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell 173, 400-416.e11. https://doi.org/10.1016/j.cell.2018.02.052

Markosyan, N., Li, J., Sun, Y.H., Richman, L.P., Lin, J.H., Yan, F., Quinones, L., Sela, Y., Yamazoe, T., Gordon, N., Tobias, J.W., Byrne, K.T., Rech, A.J., FitzGerald, G.A., Stanger, B.Z., Vonderheide, R.H., 2019. Tumor cell–intrinsic EPHA2 suppresses antitumor immunity by regulating PTGS2 (COX-2). J. Clin. Invest. 129, 3594–3609. https://doi.org/10.1172/JCI127755

Markowitz, S.D., Bertagnolli, M.M., 2009. Molecular Basis of Colorectal Cancer. N. Engl. J. Med. 361, 2449–2460. https://doi.org/10.1056/NEJMra0804588

Matsumoto, N., Kubo, A., Liu, H., Akita, K., Laub, F., Ramirez, F., Keller, G., Friedman, S.L., 2006. Developmental regulation of yolk sac hematopoiesis by Kruppel-like factor 6. Blood 107, 1357–1365. https://doi.org/10.1182/blood-2005-05-1916

McGee, R.S., Kosterlitz, O., Kaznatcheev, A., Kerr, B., Bergstrom, C.T., 2022. The cost of information acquisition by natural selection. https://doi.org/10.1101/2022.07.02.498577

Mémoli, F., 2017. On the use of Gromov-Hausdorff Distances for Shape Comparison. Eurographics Symp. Point-Based Graph. 10.

Monaco, G., Lee, B., Xu, W., Mustafah, S., Hwang, Y.Y., Carré, C., Burdin, N., Visan, L., Ceccarelli, M., Poidinger, M., Zippelius, A., Pedro de Magalhães, J., Larbi, A., 2019. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. Cell Rep. 26, 1627-1640.e7. https://doi.org/10.1016/j.celrep.2019.01.041

Moon, E.K., Wang, L.-C.S., Bekdache, K., Lynn, R.C., Lo, A., Thorne, S.H., Albelda, S.M., 2018. Intra-tumoral delivery of CXCL11 via a vaccinia virus, but not by modified T cells, enhances the efficacy of adoptive T cell therapy and vaccines. Oncoimmunology 7, e1395997. https://doi.org/10.1080/2162402X.2017.1395997

Nitzan, M., Karaiskos, N., Friedman, N., Rajewsky, N., 2019. Gene expression cartography. Nature 576, 132–137. https://doi.org/10.1038/s41586-019-1773-3

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic

expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745. https://doi.org/10.1093/nar/gkv1189

Palla, G., Fischer, D.S., Regev, A., Theis, F.J., 2022. Spatial components of molecular tissue biology. Nat. Biotechnol. 40, 308–318.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., others, 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Peyr\'e, G., Cuturi, M., 2019. Computational Optimal Transport. Found. Trends Mach. Learn. 11, 355–607.

Phillips, D., Matusiak, M., Gutierrez, B.R., Bhate, S.S., Barlow, G.L., Jiang, S., Demeter, J., Smythe, K.S., Pierce, R.H., Fling, S.P., Ramchurren, N., Cheever, M.A., Goltsev, Y., West, R.B., Khodadoust, M.S., Kim, Y.H., Schürch, C.M., Nolan, G.P., 2021. Immune cell topography predicts response to PD-1 blockade in cutaneous T cell lymphoma. Nat. Commun. 12, 6726. https://doi.org/10.1038/s41467-021-26974-6

Przybyla, L., Gilbert, L.A., 2022. A new era in functional genomics screens. Nat. Rev. Genet. 23, 89–103. https://doi.org/10.1038/s41576-021-00409-w

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., Lim, W.A., 2013. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. Cell 152, 1173–1183. https://doi.org/10.1016/j.cell.2013.02.022

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the 38th International Conference on Machine Learning. PMLR, pp. 8748–8763.

Riley, M., Abe, T., Arnaud, M.B., Berlyn, M.K.B., Blattner, F.R., Chaudhuri, R.R., Glasner, J.D., Horiuchi, T., Keseler, I.M., Kosuge, T., Mori, H., Perna, N.T., Plunkett, G., Rudd, K.E., Serres, M.H., Thomas, G.H., Thomson, N.R., Wishart, D., Wanner, B.L., 2006. Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. Nucleic Acids Res. 34, 1–9. https://doi.org/10.1093/nar/gkj405

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hickey, G., Hinrichs, A.S., Hubley, R., Karolchik, D., Learned, K., Lee, B.T., Li, C.H., Miga, K.H., Nguyen, N., Paten, B., Raney, B.J., Smit, A.F.A., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., Kent, W.J., 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. 43, D670–D681. https://doi.org/10.1093/nar/gku1177

Rossi, G., Manfrin, A., Lutolf, M.P., 2018. Progress and potential in organoid research. Nat. Rev. Genet. 19, 671–687.

Samusik, N., Good, Z., Spitzer, M.H., Davis, K.L., Nolan, G.P., 2016. Automated mapping of phenotype space with single-cell data. Nat. Methods 13, 493–6. https://doi.org/10.1038/nmeth.3863

Satija, R., Hoffman, P., Butler, A., 2019. SeuratData: Install and manage Seurat datasets. R Package 576.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., others, 2019. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell 176, 928–943.

Schmidt, R., Steinhart, Z., Layeghi, M., Freimer, J.W., Bueno, R., Nguyen, V.Q., Blaeschke, F., Ye, C.J., Marson, A., 2022. CRISPR activation and interference screens decode stimulation responses in primary human T cells. Science 375, eabj4008. https://doi.org/10.1126/science.abj4008

Schmiedel, B.J., Gonzalez-Colin, C., Fajardo, V., Rocha, J., Madrigal, A., Ramírez-Suástegui, C., Bhattacharyya, S., Simon, H., Greenbaum, J.A., Peters, B., Seumois, G., Ay, F., Chandra, V., Vijayanand, P., 2022. Single-cell eQTL analysis of activated T cell subsets reveals activation and cell type–dependent effects of disease-risk variants. Sci. Immunol. 7, eabm2508. https://doi.org/10.1126/sciimmunol.abm2508

Schmiedel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J.A., McVicker, G., Seumois, G., Rao, A., Kronenberg, M., Peters, B., Vijayanand, P., 2018. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell 175, 1701-1715.e16. https://doi.org/10.1016/j.cell.2018.10.022

Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., Fulton, R.S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K.M., Auger, K., Chow, W., Collins, J., Harden, G., Hubbard, T., Pelan, S., Simpson, J.T., Threadgold, G., Torrance, J., Wood, J.M., Clarke, L., Koren, S., Boitano, M., Peluso, P., Li, H., Chin, C.-S., Phillippy, A.M., Durbin, R., Wilson, R.K., Flicek, P., Eichler, E.E., Church, D.M., 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864. https://doi.org/10.1101/gr.213611.116

Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y., 2021. Toward causal representation learning. Proc. IEEE 109, 612–634.

Schürch, C.M., Bhate, S.S., Barlow, G.L., Phillips, D.J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D.R., Kinoshita, S., Samusik, N., Goltsev, Y., Nolan, G.P., 2020. Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. Cell 182, 1341-1359.e19. https://doi.org/10.1016/j.cell.2020.07.005

Severin, I.C., Gaudry, J.-P., Johnson, Z., Kungl, A., Jansma, A., Gesslbauer, B., Mulloy, B., Power, C., Proudfoot, A.E.I., Handel, T., 2010. Characterization of the Chemokine CXCL11-Heparin Interaction Suggests Two Different Affinities for Glycosaminoglycans. J. Biol. Chem. 285, 17713–17724. https://doi.org/10.1074/jbc.M109.082552

Sun, G., Cao, Y., Qian, C., Wan, Z., Zhu, J., Guo, J., Shi, L., 2020. Romo1 is involved in the immune response of glioblastoma by regulating the function of macrophages. Aging 12, 1114–1127. https://doi.org/10.18632/aging.102648

Uhlen, M., Karlsson, M.J., Zhong, W., Tebani, A., Pou, C., Mikes, J., Lakshmikanth, T., Forsström, B., Edfors, F., Odeberg, J., Mardinoglu, A., Zhang, C., von Feilitzen, K., Mulder, J., Sjöstedt, E., Hober, A., Oksvold, P., Zwahlen, M., Ponten, F., Lindskog, C., Sivertsson, Å.,

Fagerberg, L., Brodin, P., 2019. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science 366, eaax9198. https://doi.org/10.1126/science.aax9198

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., Sanli, K., von Feilitzen, K., Oksvold, P., Lundberg, E., Hober, S., Nilsson, P., Mattsson, J., Schwenk, J.M., Brunnström, H., Glimelius, B., Sjöblom, T., Edqvist, P.-H., Djureinovic, D., Micke, P., Lindskog, C., Mardinoglu, A., Ponten, F., 2017. A pathology atlas of the human cancer transcriptome. Science 357, eaan2507. https://doi.org/10.1126/science.aan2507

Vaishnav, E.D., de Boer, C.G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., Thompson, D.A., Levin, J.Z., Cubillos, F.A., Regev, A., 2022. The evolution, evolvability and engineering of gene regulatory DNA. Nature 603, 455–463. https://doi.org/10.1038/s41586-022-04506-6

van der Wijst, M., de Vries, D., Groot, H., Trynka, G., Hon, C., Bonder, M., Stegle, O., Nawijn, M., Idaghdour, Y., van der Harst, P., Ye, C., Powell, J., Theis, F., Mahfouz, A., Heinig, M., Franke, L., 2020. The single-cell eQTLGen consortium. eLife 9, e52155. https://doi.org/10.7554/eLife.52155

Vollmer, T., Schlickeiser, S., Amini, L., Schulenberg, S., Wendering, D.J., Banday, V., Jurisch, A., Noster, R., Kunkel, D., Brindle, N.R., Savidis, I., Akyüz, L., Hecht, J., Stervbo, U., Roch, T., Babel, N., Reinke, P., Winqvist, O., Sherif, A., Volk, H.-D., Schmueck-Henneresse, M., 2021. The intratumoral CXCR3 chemokine system is predictive of chemotherapy response in human bladder cancer. Sci. Transl. Med. 13, eabb3735. https://doi.org/10.1126/scitranslmed.abb3735

Wagner, G.P., Zhang, J., 2011. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nat. Rev. Genet. 12, 204–213. https://doi.org/10.1038/nrg2949

Walker, V.M., Zheng, J., Gaunt, T.R., Smith, G.D., 2022. Phenotypic Causal Inference Using Genome-Wide Association Study Data: Mendelian Randomization and Beyond. Annu. Rev. Biomed. Data Sci. 5, null. https://doi.org/10.1146/annurev-biodatasci-122120-024910

Wang, J., Zhang, X., Li, J., Ma, X., Feng, F., Liu, L., Wu, J., Sun, C., 2020. ADRB1 was identified as a potential biomarker for breast cancer by the co-analysis of tumor mutational burden and immune infiltration. Aging 13, 351–363. https://doi.org/10.18632/aging.104204

Wangler, M.F., Yamamoto, S., Chao, H.-T., Posey, J.E., Westerfield, M., Postlethwait, J., Members of the Undiagnosed Diseases Network (UDN), Hieter, P., Boycott, K.M., Campeau, P.M., Bellen, H.J., 2017. Model Organisms Facilitate Rare Disease Diagnosis and Therapeutic Research. Genetics 207, 9–27. https://doi.org/10.1534/genetics.117.203067

Watson, H.A., Durairaj, R.R.P., Ohme, J., Alatsatianos, M., Almutairi, H., Mohammed, R.N., Vigar, M., Reed, S.G., Paisey, S.J., Marshall, C., Gallimore, A., Ager, A., 2019. L-Selectin Enhanced T Cells Improve the Efficacy of Cancer Immunotherapy. Front. Immunol. 10.

Wolf, F.A., Angerer, P., Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15. https://doi.org/10.1186/s13059-017-1382-0

Xiong, Z., Ampudia Mesias, E., Pluhar, G.E., Rathe, S.K., Largaespada, D.A., Sham, Y.Y., Moertel, C.L., Olin, M.R., 2020. CD200 Checkpoint Reversal: A Novel Approach to Immunotherapy. Clin. Cancer Res. 26, 232–241. https://doi.org/10.1158/1078-0432.CCR-19-2234

Yang, K.D., Belyaeva, A., Venkatachalapathy, S., Damodaran, K., Katcoff, A., Radhakrishnan, A., Shivashankar, G.V., Uhler, C., 2021. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. Nat. Commun. 12, 31. https://doi.org/10.1038/s41467-020-20249-2

Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M.G., Andersen, S., Lu, Q., Rowson, A., Taylor, T.R.P., Clarke, L., Maccora, K., Chen, C., Cook, A.L., Ye, C.J., Fairfax, K.A., Hewitt, A.W., Powell, J.E., 2022. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. Science 376, eabf3041. https://doi.org/10.1126/science.abf3041

Zhou, J., Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods 12, 931–934. https://doi.org/10.1038/nmeth.3547

Figure Legends

Figure 1. Defining templates of a CMP using causal phenotype sequence alignment

A. Conceptual summary of theory. 1) The genome (left, represented by black box) produces complex molecular phenotypes (CMPs), represented by blue structure. 2) A template of the CMP can be found within the genome (black structure of similar shape to CMP). 3) Given any edit to CMP (original CMP on left and edited CMP on right, top row), correspondingly editing the template (bottom row, template in genome resembles new CMP) results in production of the correctly edited CMP (structure on bottom row matches top row correctly, indicated by green tick).

B. Schematic illustration of the causal phenotype sequence alignment procedure. 1) An observed CMP (left) and the genome are represented as metric spaces (with black points and distances as seen visually on the page) 2) A phenotype sequence alignment (PSA) is a map sending points in the CMP to points the genome while preserving distances. Black dashed arrows indicate where points in phenotype space are sent in genome. 3) Editing the CMP is modelled as moving selected points of the CMP in selected directions, visualized by blue arrows. 4) Causal phenotype sequence alignment (CPSA) maps the edit from 3) to the genome. The alignment is causal if genetically manipulating the genomic loci corresponding to the edit results in production of the correctly edited CMP.

C. Three ingredients in formal definition of CMP. CMPs are datapoints in a metric space referred to as phenotype space. Left: each point in phenotype space is a possible observation vector, assigning a real number to each feature (examples listed above). Top right: The distance between points in phenotype space is the standard Euclidean distance. Lower right: An observation of a CMP is a collection of points in phenotype space, as visualized in the oval.

D. Representation of a single genome as a metric space. Left: the points in the metric space are k-mer functionals, two examples shown in red and green boxes. Each k-mer functional assigns a real number to every k-mer in the genome. Right: the distance between two functionals (red and green lines respectively) in the metric space are obtained by summing the squared differences in their values on each k-mer in the genome (orange differences between lines). The distances between two functionals depends on how often each k-mer appears, so the metric encodes the sequence of the genome.

E. Visualizing phenotype sequence alignments of CMPs with the genome. Left: the CMP is a collection of points in phenotype space as in **Figure 1C**. Each pair of points has a distance (colored arrows between points). The alignment T sends points in phenotype space to k-mer functionals in the genome in a distance preserving way. Right: The areas (labelled a,b,c,d,e,f) between the curves $T(x_1)$, $T(x_2)$, $T(x_3)$ (red, green and blue respectively) must satisfy the equations listed if T preserve distances.

Figure 2. PSAs capture specific genetic information from a single phenotypic measurement

A. Summary of the Neural Alignment of CMP and Reference genome (NACR) algorithm. Left, top: the input data is only: a CMP and reference genome sequence (no genetic variation, biological annotations or external data are used). Left, bottom: the output is a function T mapping points in phenotype space to k-mer functionals while preserving distances. Right: Two neural networks F and G extract m-dimensional representations of datapoints and k-mers respectively. The value of functional T(x) on k-mer s is the inner product of these representations, i.e. $T(x)(s) = \langle F(x), G(s) \rangle$.

B. Summary of applying NACR to find PSA of high-parameter tissue image with a reference genome. Image is represented as a of cell type counts (identified by segmentation, with no biological annotations) within image patches. 128-mers are extracted from the reference genome. An "alignment loss" trains F and G as in Figure 2A to enforce that T is a PSA. Regularization terms are included to restrict complexity of PSA.

C. Quality of PSAs of human tonsil with different genomes as permitted complexity of NACR is varied. Each plotted point indicates the alignment loss (y axis) obtained by training NACR from initialization on a set of 10000 128-mers extracted from the reference genome of the species indicated by the point's color as the alignment complexity (x axis) is restricted by L2 regularization on neural network weights, see Methods. ** indicates Mann Whitney $p < 0.01$, ***= $p< 0.001$.

D. PSA scores genes by correspondence with points in phenotype space. Given a point x in phenotype space and an alignment T, the functional T(x) is aggregated over genes by summing 128-mers extracted between the transcription start site and transcription end site to compute a gene functional corresponding to x.

E. PSA of human tonsil with GRCh38 highlights tonsil-specific gene modules. Table of significantly enriched gene sets when genes are ranked by the difference in correspondence with Dark-Zone

and Light-Zone in PSA from applying NACR to align tonsil CODEX data with GRCh38. GSEA FDR indicates the Benjamini-Hochberg procedure adjusted p-values. Train FDR indicates the p-value relative to null alignments (i.e., without training NACR) in which that GO term was also significantly enriched, adjusted for multiple hypotheses.

F. PSA of CRC immune tumor microenvironment with GRCh38 highlights CRC-specific gene modules. Table of significantly enriched gene sets when genes are ranked by the difference in correspondence with CLR and DII patients in alignment of CRC CODEX data with GRCh38. Columns as **Figure 2E**.

G. Association of functional values with pathogenic SNPs. The value of the functional in **Figure 2F** was evaluated on 128-mers centered at each of 419 ClinVar single-nucleotide polymorphisms that were expert-validated as pathogenic and contained the search term 'colorectal' (right hand violin and points) and 10000 128-mers centered on positions extracted at random within the 1MB up-and downstream of the SNPs were evaluated. Train p indicates the frequency of null alignments with a less than or equal permutation p-value.

H. PSAs preferentially highlight different types of genomic region. 20000 random 128-mers from promoter (2kb upstream of TSS), terminator (2kb downstream of transcription end site), transcript and intergenic (not promoter, transcript or terminator) regions were extracted from each chromosome. Average absolute values for the functional used for **Figure 2E** (top) and **2F** (bottom) for each chromosome are plotted. Significant p-values in **Supplementary Figure 1A**.

Figure 3. PSAs of single-cell gene expression distributions are templates

A. Implenting causal phenotype sequence alignment with discrepancy functionals. Top left: The original CMP is represented as clusters in phenotype space (green circles, size indicate size of cluster). A phenotypic edit is modelled as shifting cluster means (x_1, x_2, x_3) to (x_1 +dx_1, x_2 +dx_2, x_3 +dx_3) in phenotypic space. Blue arrow shifts green circle at x_1 to red circle at x_1 + dx_1; dx_2 = dx_3 = 0. Discrepancy functional (blue trace in lower right) is obtained by taking the squared differences (dashed blue lines) at each k-mer between the functional assigned to T(x_1) and T(x_1 +dx_1) (top right, red and green traces respectively). Lower left: formula for the discrepancy functional. The PSA a template if genomic loci with high discrepancy are those whose experimental manipulation achieve the edit.

B. Simulation strategy to assess whether information-efficiency of the genome imposes CMPs to have templates. Top: a random binary genome is sampled and a neural network Φ is trained

with one sample (the sampled genome) to output a mixture of two Gaussians (centered at -1 and 1) with or without regularization (information-efficiency). This neural network is the genotype-phenotype map. Lower, left column, top: a PSA is identified by NACR of the output genome with the sampled CMP. Lower, left column, bottom: the edit to the CMP is a shift in one of the mixture components (b -> b'). Lower, middle column: T(b) and T(b') are computed at each 3-mer yielding a discrepancy $(T(b)-T(b'))^2$. Right column: each bit in the genome is flipped one-by-one and the resultant CMP is computed by applying Φ, and the distance to the target phenotype (Wasserstein distance) is measured. The correlation between the discrepancies and the distance to target CMP are assessed.

C. Information-efficiency of the genotype-phenotype map enforces CMPs to have templates in simulations. Spearman correlation between discrepancies and distance to target CMP across 20 binary genome samples under varying regularization of the genomic model (x axis) and varying information-efficiency constraints (color, regularization of Φ) using the simulation strategy in **Figure 3B**. * = t-test $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$. Blue * indicates $p< 0.05$ when comparing absolute values of blue "untrained" and "unregularized" violins. Orange ** indicates $p<0.01$ when comparing absolute values of orange "untrained" and "unregularized" violins.

D. Overview of applying CPSA to find genomic loci that edit cell type specific gene expression. Top: an alignment T of a single-cell gene expression dataset with a reference genome is found by PSA. Bottom: the discrepancy between the functionals assigned to the original average gene expression vector of a cell type and the shifted gene expression vector are computed at each k-mer, as in **Figure 2A.**

E. Discrepancies for editing cell type specific gene expression are consistent across biological replicates. The target edits were increasing values of gene expression in direction of top three principal components monocytes, $CD4^+$ T cells and $CD8^+$ T cells (x axis). Discrepancy functionals for edits were obtained from PSAs of PBMC single-cell gene expression of individual donors in (Hao et al., 2021) dataset with GRCh38, with either no training or varying regularization (color) were. Points indicate Spearman correlations between discrepancies evaluated on 50000 random k-mers from each pair of donors PSAs.

F. Discrepancies for increasing IFNG expression in $CD8^+$ T cells correspond to gene-hits in CRISPR activation screen for increasing interferon gamma protein expression. Discrepancies for increasing IFNG expression in $CD8^+$ T cells were computed using an alignment of PBMC single-cell gene expression dataset with GRCh38 and subsequently aggregated over genes as per

Figure 2D and normalized. Receiver-operator-characteristic curve for using normalized gene-level discrepancies to select gene hits in CRISPR activation screen for interferon-gamma expression. Permutation p-value for area-under-curve (AUC) reported.

G. Association with CRISPR screen hits for interferon gamma expression is specific to discrepancy for increasing IFNG expression. Normalized gene level discrepancies for increasing each of 223 genes with similar expression patterns to IFNG in $CD8^+$ T cells in PBMC dataset were evaluated for AUC (x axis)/p-values (y axis) predicting CRISPR screen hits. Black points indicate permutation genes p-value < 0.05 (horizontal line). Red star is IFNG.

H. Discrepancies for shifting $CD4^+$ T cell effectorness genes reveal regulatory modules. Normalized discrepancies for each gene (rows) for increasing expression of $CD4^+$ T cell "effectorness" genes one-by-one (columns). Color of heatmap indicates normalized discrepancy. Union of 200 most discrepant genes (rows) for each effectorness gene (column) shown. Row color: hierarchical clustering and cutting with a maximum of 15 clusters (row colors). Below: enriched GO biological processes (FDR < 0.1 in individual clusters) and corresponding genes in cluster.

I. Discrepancies for editing organogenesis lineage composition reveal lineage-specific master transcription factors. Normalized discrepancies of genes (rows) corresponding to increasing the frequency of cell lineages (columns) in alignment of mouse organogenesis single-cell gene expression atlas dataset with mm10. Union of 200 most discrepant genes (rows) for each developmental lineage (column) shown. Row color: hierarchical clustering and cutting with a maximum of 15 clusters (row colors). Below: enriched GO biological processes (FDR < 0.1 in individual clusters) and corresponding genes in cluster.

Figure 4. PSAs of the CRC iTME with GRCh38 are templates

A. Illustration of target edit to CMP. Target edit is to increase the frequency of CD8+ T cells (black dots) in the tumor cellular neighborhood (CN, green).

B. Discrepancies for target edit are consistent across PSAs obtained from multiple NACR training runs from random initializations. Spearman correlation in gene-level discrepancies for target edit compared between pairs of PSAs (lower) from null alignments (randomly initialized but not trained) with the same architecture. Input to NACR: CRC iTME CODEX data and GRch38. 1-sided permutation p-value.

C. Discrepancies reveal genes associated with editing tumor infiltrating $CD8^+$ T cells. Top twenty genes by discrepancy increase in 6 PSAs relative to 99 null alignments (t-scores, x axis) are

plotted by association with CD8A expression in tumors from TCGA COAD-READ dataset (y-axis). Red point (CXCL11) indicates FDR < 0.1 and significant association with CD8A expression. 15/20 were significantly associated with CD8A expression. P values: 3.7% subsets of null alignments had 15 or more of top 20 genes associated with CD8A expression; 53% genes significantly associated, n = 20 genes, Binomial p-value = 0.034. PSAs as **Figure 4B.**

D. Discrepancies identify an active site in CXCL11 whose mutation alters lymphocyte trafficking. Permutation p-value for increase in discrepancy in PSAs vs null alignments for target edit evaluated at 128-mers at centered at each position in transcript sequence of CXCL11 (log p value on y axis) at each position. Exon and coding sequence boundaries as indicated. Black points indicate permutation p< 0.01. Lower: the amino acid translation of exon 2 is illustrated colored by p-value for discrepancy increase in PSAs. The position with significantly higher discrepancy was previously shown in mutagenesis studies to affect trafficking of lymphocytes. PSAs as **Figure 4B**.

E. Discrepancies for target edit identify sequences of gene modules associated with antitumoral immune response. Significantly enriched GO biological process categories, genes and FDR amongst the top 100 genes with highest discrepancy for edit obtained from PSA of CRC dataset with GRch38 in Figure 2.

F. Mutations in discrepant genes associated with patient survival. Survival analysis in TCGA data (image credit cBioHub) comparing patients with alterations in CD200R1, MMP3 or PTGS2 to patients without.

G. Genetic manipulation of discrepant gene edits tumor infiltrating $CD8^+$ T cells. Previously reported data under CCBY 4.0 license from (Sun et al., 2020) showing lentiviral overexpression of ROMO1 in transplanted bone-marrow cells reduces intratumoral T cells in a mouse model of glioblastoma.

H. Genetic manipulation of discrepant gene edits tumor infiltrating CD8 T cells. Previously reported data under CCBY 4.0 license from (Markosyan et al., 2019) showing PTGS2 KO in pancreatic tumors increases intratumoral CD8+ T cell frequency.

I. Discrepancies identify promoter regions associated with $CD8^+$ T cell trafficking. Enriched gene modules amongst top 100 genes ranked by discrepancy for target edit aggregated over their promoter regions (5kb upstream of TSS).

J. Target edit: large-scale architecture change, edit the extrafollicular regions of the iTME from resembling that of diffuse immune infiltrate (DII) patients to a Crohn's like reaction (CLR)

patients. Optimal transport coupling (matrix on right) was computed between DII patients' and CLR patients' samples and used to compute discrepancies for architecture change.

K. Enriched GO Biological Processes amongst top 100 most discrepant genes for architecture change edit. Underlined genes indicate those previously shown to be associated with enhanced B cell infiltrate in cancer. CXCR4 is bold because its pharmacological inhibition was shown to induce a B cell response in CRC patients.

Supplementary Figure Legends

Supplementary Figure 1

A. Significant (permutation p-value < 0.05) differences between absolute functional values in Figure 2H between 128-mers extracted from different types of sequence region.

Supplementary Figure 2.

A. Spearman correlations (y axis) between discrepancies for increasing different PCs in different cell types (x axis) pairs of patients at sparsity and complexity regularization hyperparameter = 0.1.

B. Heatmap indicating discrepancies for increasing gene (column) aggregated over genes in genome (row), z-normalized across rows.

C. Heatmap after the first singular value had been removed from matrix in **B.**

D. -log10(permutation p-value) (y-axis) for difference in median normalized discrepancy for increasing IFNG expression between hits and non-hits (blue) or AUC for using high discrepancy for increasing IFNG expression to indicate hits (orange) as absolute z-score threshold for defining hits in CRISPR screen is varied (x-axis).

E. Observed difference in median normalized discrepancy between hits and non-hits in CRISPR screen (x-axis) corresponding to shifting mean $CD8^+$ T cell expression in direction of each of 223 genes with similar expression patterns as IFNG. Each point is a gene in the genome over which the discrepancies are aggregated.

F. Pearson correlation coefficient between IFNG expression and normalized discrepancy for increasing IFNG. Each point is a gene in the genome over which the discrepancies are aggregated.

G. Histogram of p-values when discrepancies are normalized (removal of first singular value) using 20 samples of genes for whose shift the discrepancy is computed.

H. Table of top 5 genes (left column) by normalized discrepancy (right column) for increasing IFNG expression

I. Extending CPSA to find discrepancies in **Figure 3A, lower left** for general phenotypic shifts using optimal transport between original and target phenotype (image). Expression for discrepancy (formula) (see **Supplementary Note 1, Section 3**).

Supplementary Figure 3.

A. Empirical null distribution of % of top 20 hits that have significant association with CD8A expression when subsets of 6 null alignments are compared to other 93 null alignments (across 5000 samples). Line indicates 0.75, which is test statistic observed when comparing 6 PSAs to 99 null alignments. Area on right of line is 3.7% of models.

B. Top genes by normalized discrepancy increase in PSAs relative to null alignments for shifting iTME organization from DII-like to CLR-like.

C. Kaplan-Meier overall survival curves for high expression vs. low expression of SLC30A1 in Stage iv, iva and ivb CRC patients (best expression cut off). TCGA data, image credit: Human Protein Atlas.

Figure 1

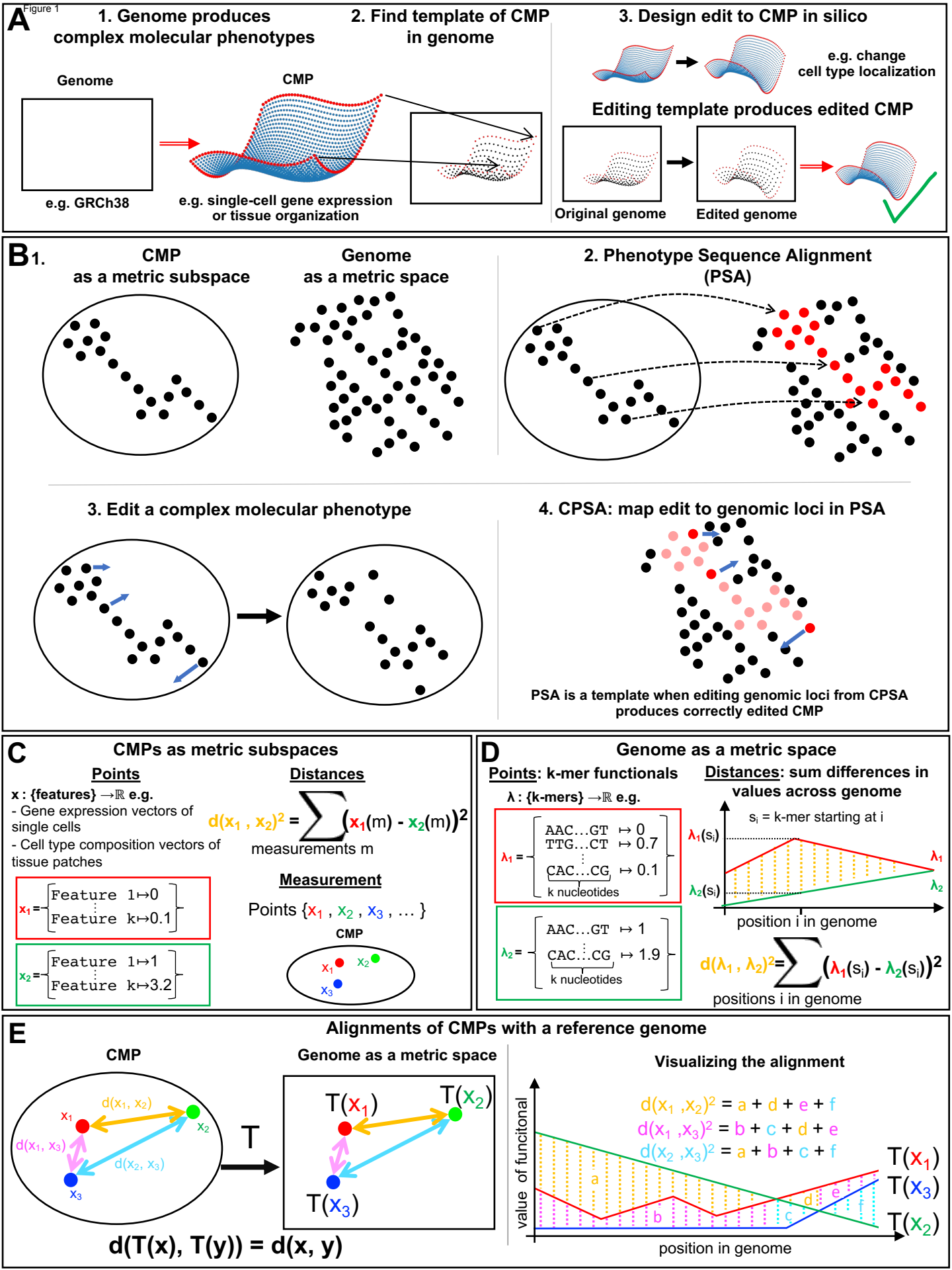

A
1. Genome produces complex molecular phenotypes
2. Find template of CMP in genome
3. Design edit to CMP in silico
Genome
CMP
e.g. GRCh38
e.g. single-cell gene expression or tissue organization
e.g. change cell type localization
Editing template produces edited CMP
Original genome
Edited genome
B
1.
CMP as a metric subspace
Genome as a metric space
2. Phenotype Sequence Alignment (PSA)
3. Edit a complex molecular phenotype
4. CPSA: map edit to genomic loci in PSA
PSA is a template when editing genomic loci from CPSA produces correctly edited CMP
C
CMPs as metric subspaces
Points
x : {features} →ℝ e.g.
- Gene expression vectors of single cells
- Cell type composition vectors of tissue patches
Feature 1↦0
Feature k↦0.1
Feature 1↦1
Feature k↦3.2
Distances
d(x1 , x2)2 = ∑ (x1(m) - x2(m))2
measurements m
Measurement
Points {x1 , x2 , x3 , … }
CMP
D
Genome as a metric space
Points: k-mer functionals
λ : {k-mers} →ℝ e.g.
AAC…GT ↦ 0
TTG…CT ↦ 0.7
CAC…CG ↦ 0.1
k nucleotides
AAC…GT ↦ 1
CAC…CG ↦ 1.9
k nucleotides
Distances: sum differences in values across genome
si = k-mer starting at i
λ1(si)
λ2(si)
position i in genome
d(λ1 , λ2)2= ∑ (λ1(si) - λ2(si))2
positions i in genome
E
Alignments of CMPs with a reference genome
CMP
Genome as a metric space
d(x1, x2)
d(x1, x3)
d(x2, x3)
T
T(x1)
T(x2)
T(x3)
d(T(x), T(y)) = d(x, y)
Visualizing the alignment
d(x1 ,x2)2 = a + d + e + f
d(x1 ,x3)2 = b + c + d + e
d(x2 ,x3)2 = a + b + c + f
value of funcitonal
position in genome

Figure 2

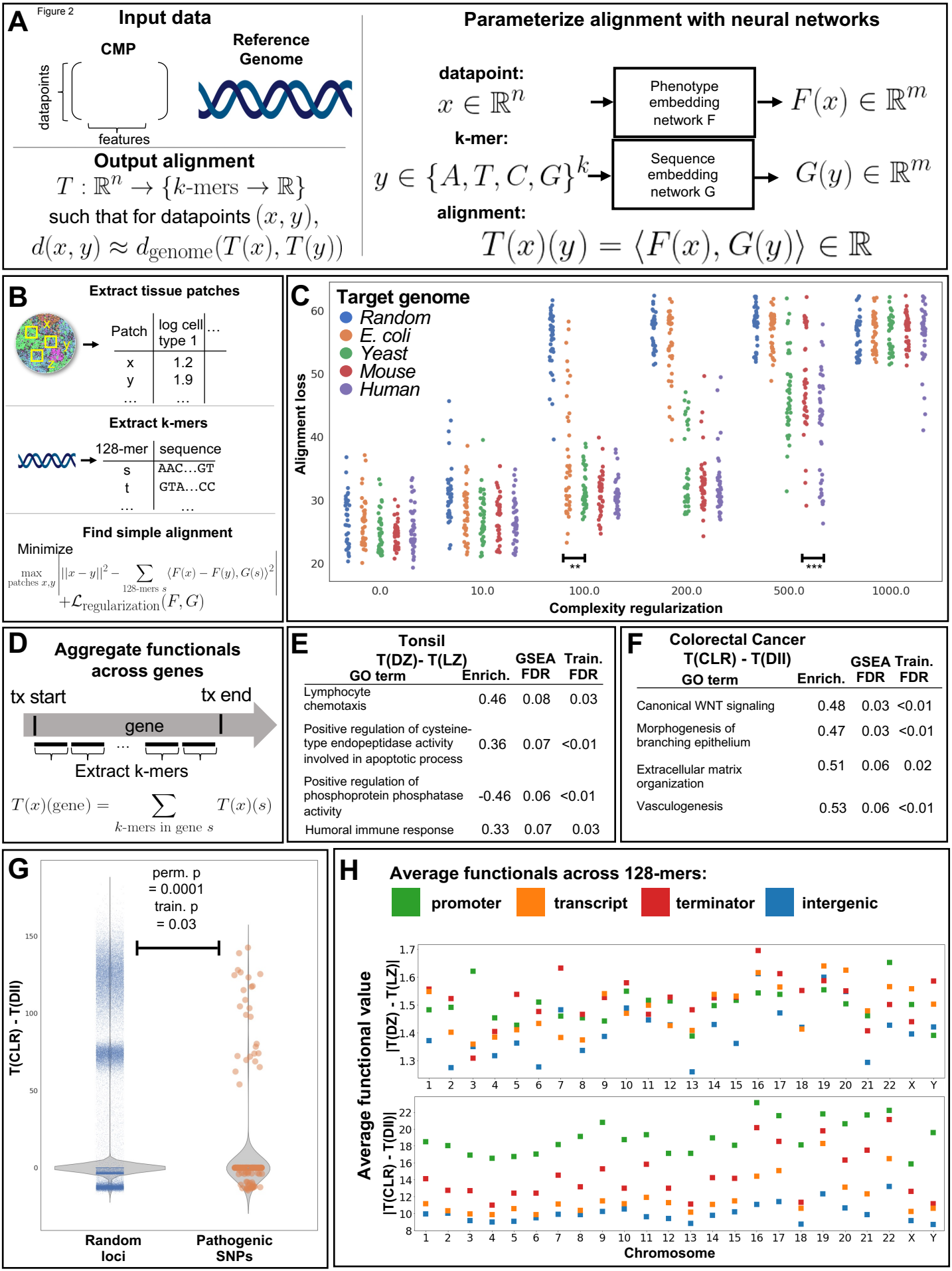

A
Input data
CMP
datapoints
features
Reference Genome
Output alignment
$T : \mathbb{R}^n \to \{k\text{-mers} \to \mathbb{R}\}$
such that for datapoints (x, y),
$d(x, y) \approx d_{\text{genome}}(T(x), T(y))$
Parameterize alignment with neural networks
datapoint:
$x \in \mathbb{R}^n$
k-mer:
$y \in \{A, T, C, G\}^k$
Phenotype embedding network F
Sequence embedding network G
$F(x) \in \mathbb{R}^m$
$G(y) \in \mathbb{R}^m$
alignment:
$T(x)(y) = \langle F(x), G(y) \rangle \in \mathbb{R}$
B
Extract tissue patches
Patch
log cell type 1
x
y
1.2
1.9
Extract k-mers
128-mer
sequence
s
AAC…GT
t
GTA…CC
Find simple alignment
Minimize
$\max_{\text{patches } x,y} \left| ||x - y||^2 - \sum_{\text{128-mers } s} \langle F(x) - F(y), G(s) \rangle^2 \right| + \mathcal{L}_{\text{regularization}}(F, G)$
C
Target genome
Random
E. coli
Yeast
Mouse
Human
Alignment loss
Complexity regularization
0.0
10.0
100.0
200.0
500.0
1000.0
D
Aggregate functionals across genes
tx start
tx end
gene
Extract k-mers
$T(x)(\text{gene}) = \sum_{k\text{-mers in gene } s} T(x)(s)$
E
Tonsil
T(DZ)- T(LZ)
GO term
Enrich.
GSEA FDR
Train. FDR
Lymphocyte chemotaxis
0.46
0.08
0.03
Positive regulation of cysteine-type endopeptidase activity involved in apoptotic process
0.36
0.07
<0.01
Positive regulation of phosphoprotein phosphatase activity
-0.46
0.06
<0.01
Humoral immune response
0.33
0.07
0.03
F
Colorectal Cancer
T(CLR) - T(DII)
GO term
Enrich.
GSEA FDR
Train. FDR
Canonical WNT signaling
0.48
0.03
<0.01
Morphogenesis of branching epithelium
0.47
0.03
<0.01
Extracellular matrix organization
0.51
0.06
0.02
Vasculogenesis
0.53
0.06
<0.01
G
perm. p = 0.0001
train. p = 0.03
T(CLR) - T(DII)
Random loci
Pathogenic SNPs
H
Average functionals across 128-mers:
promoter
transcript
terminator
intergenic
Average functional value
|T(DZ) - T(LZ)|
|T(CLR) - T(DII)|
Chromosome

Figure 3

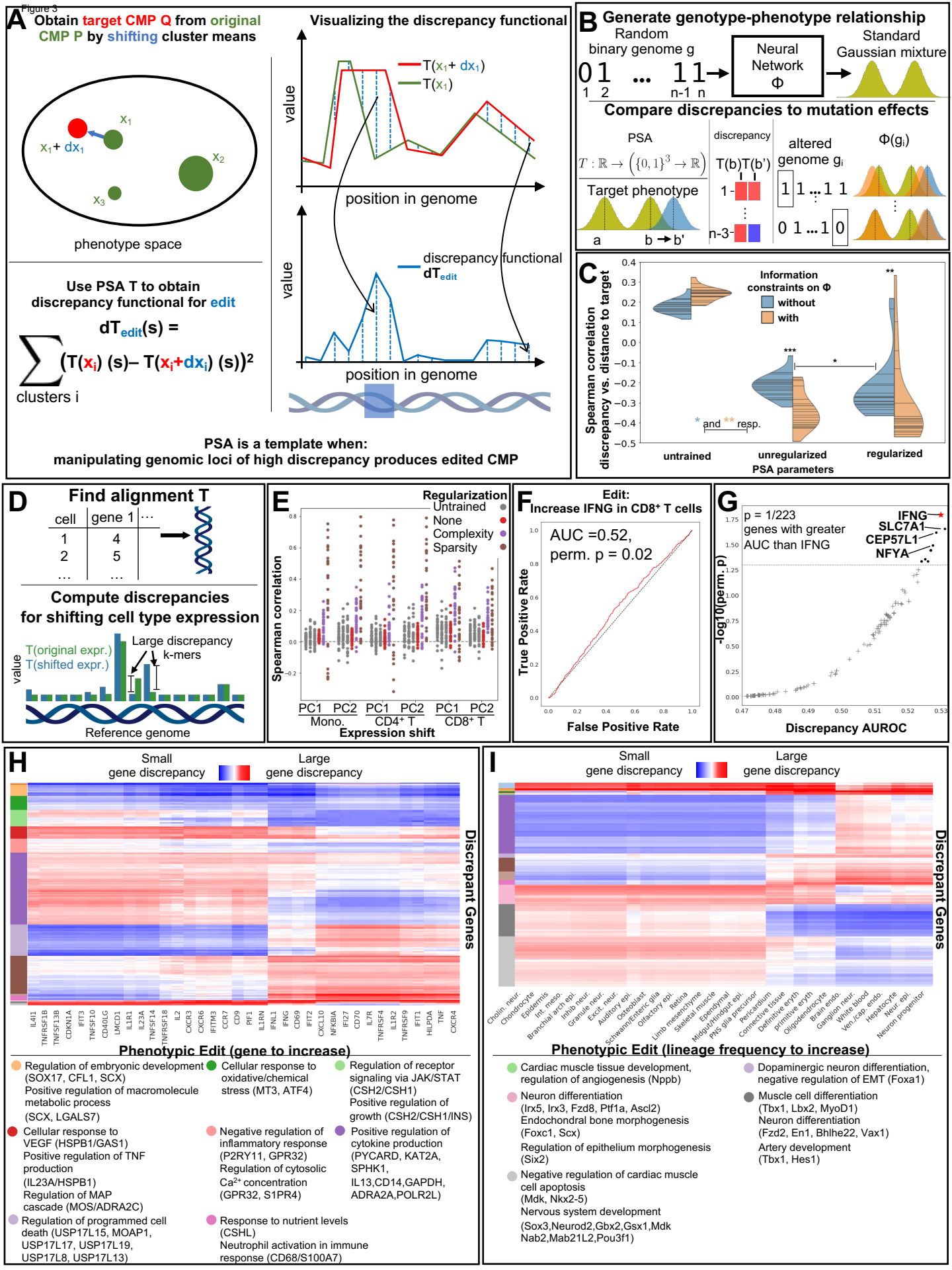
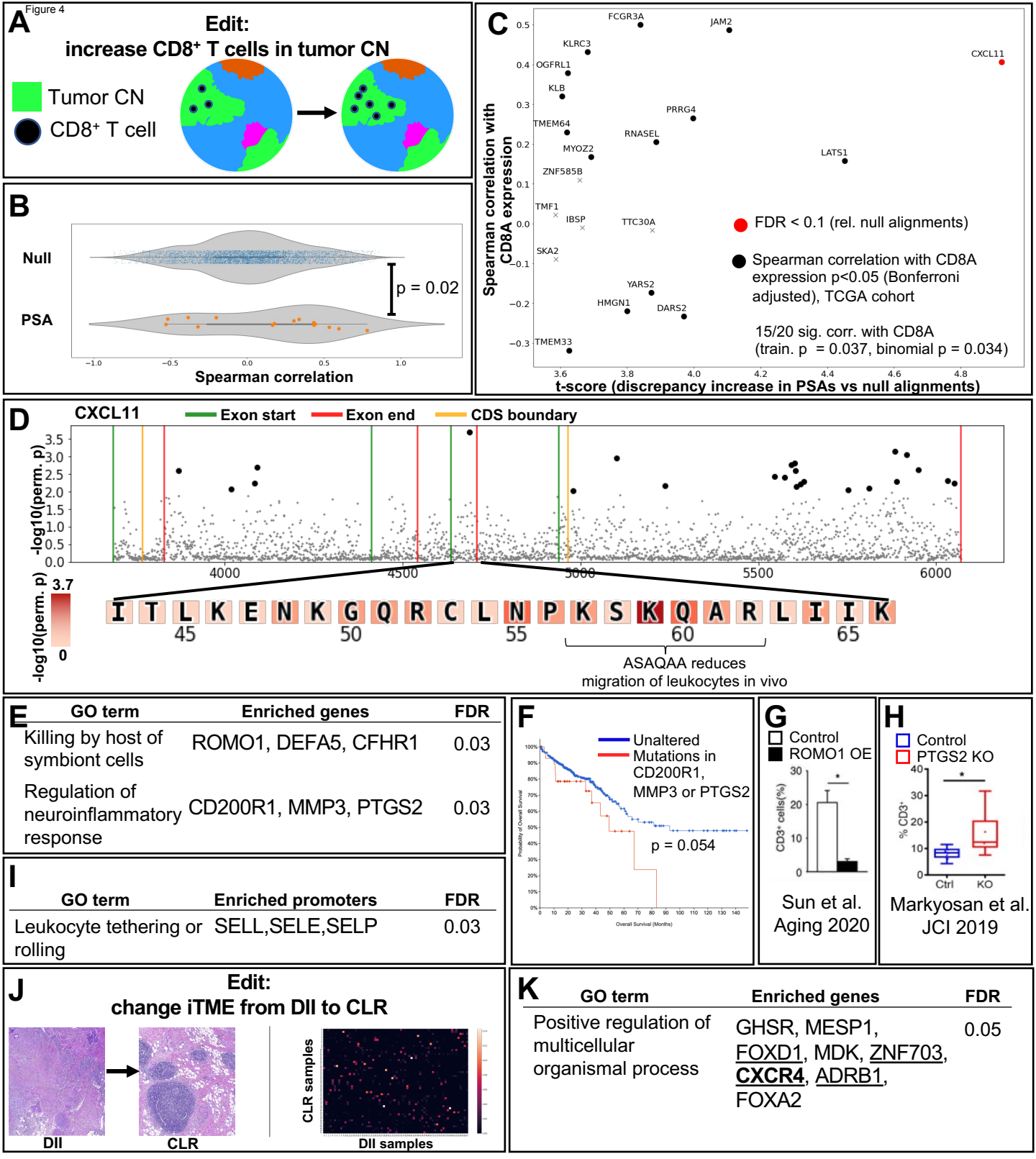

A
Obtain target CMP Q from original CMP P by shifting cluster means
x1
x1+ dx1
x2
x3
phenotype space
Use PSA T to obtain discrepancy functional for edit
dTedit(s) =
(T(xi) (s)– T(xi+dxi) (s))2
clusters i
Visualizing the discrepancy functional
T(x1+ dx1)
T(x1)
value
position in genome
discrepancy functional dTedit
value
position in genome
PSA is a template when:
manipulating genomic loci of high discrepancy produces edited CMP
B
Generate genotype-phenotype relationship
Random binary genome g
0 1 ... 1 1
1 2 n-1 n
Neural Network Φ
Standard Gaussian mixture
Compare discrepancies to mutation effects
PSA
Target phenotype
a
b → b'
discrepancy
T(b)T(b')
altered genome gi
1 1 ... 1 1
0 1 ... 1 0
Φ(gi)
C
Spearman correlation discrepancy vs. distance to target
Information constraints on Φ
without
with
*** and ** resp.
untrained
unregularized
regularized
PSA parameters
D
Find alignment T
cell
gene 1
Compute discrepancies for shifting cell type expression
T(original expr.)
T(shifted expr.)
Large discrepancy k-mers
value
Reference genome
E
Regularization
Untrained
None
Complexity
Sparsity
Spearman correlation
PC1 PC2 PC1 PC2 PC1 PC2
Mono. CD4+ T CD8+ T
Expression shift
F
Edit: Increase IFNG in CD8+ T cells
AUC =0.52,
perm. p = 0.02
True Positive Rate
False Positive Rate
G
p = 1/223 genes with greater AUC than IFNG
IFNG
SLC7A1
CEP57L1
NFYA
-log10(perm. p)
Discrepancy AUROC
H
Small gene discrepancy
Large gene discrepancy
Discrepant Genes
IL4I1 TNFRSF1B TNFSF13B CDKN1A IFIT3 TNFSF10 CD40LG LMCD1 IL1R1 IL23A TNFSF14 TNFRSF18 IL2 CXCR3 CXCR6 IFITM3 CCR7 CD9 PIF1 IL1RN IFNL1 IFNG CD69 IFIT2 CXCL10 NFKBIA IFI27 CD70 IL7R TNFRSF4 IL1R2 TNFRSF9 IFIT1 HILPDA TNF CXCR4
Phenotypic Edit (gene to increase)
Regulation of embryonic development (SOX17, CFL1, SCX)
Positive regulation of macromolecule metabolic process (SCX, LGALS7)
Cellular response to oxidative/chemical stress (MT3, ATF4)
Regulation of receptor signaling via JAK/STAT (CSH2/CSH1)
Positive regulation of growth (CSH2/CSH1/INS)
Cellular response to VEGF (HSPB1/GAS1)
Positive regulation of TNF production (IL23A/HSPB1)
Regulation of MAP cascade (MOS/ADRA2C)
Negative regulation of inflammatory response (P2RY11, GPR32)
Regulation of cytosolic Ca2+ concentration (GPR32, S1PR4)
Positive regulation of cytokine production (PYCARD, KAT2A, SPHK1, IL13,CD14,GAPDH, ADRA2A,POLR2L)
Regulation of programmed cell death (USP17L15, MOAP1, USP17L17, USP17L19, USP17L8, USP17L13)
Response to nutrient levels (CSHL)
Neutrophil activation in immune response (CD68/S100A7)
I
Small gene discrepancy
Large gene discrepancy
Discrepant Genes
Cholin. neur. Chondrocyte Epidermis Int. meso. Branchial arch epi. Inhib neur. Granule neur. Excit. neur. Auditory epi. Osteoblast Schwann/Enteric glia Olfactory epi. Retina Limb mesenchyme Skeletal muscle Ependymal Midgut/Hindgut epi. PNS glia precursor Pericardium Connective tissue Definitive eryth primitive eryth Oligodendrocyte Brain endo. Ganglion neur. White blood Ven./cap. endo. Hepatocyte Neur. epi. Neuron progenitor
Phenotypic Edit (lineage frequency to increase)
Cardiac muscle tissue development, regulation of angiogenesis (Nppb)
Neuron differentiation (Irx5, Irx3, Fzd8, Ptf1a, Ascl2)
Endochondral bone morphogenesis (Foxc1, Scx)
Regulation of epithelium morphogenesis (Six2)
Negative regulation of cardiac muscle cell apoptosis (Mdk, Nkx2-5)
Nervous system development (Sox3,Neurod2,Gbx2,Gsx1,Mdk Nab2,Mab21L2,Pou3f1)
Dopaminergic neuron differentiation, negative regulation of EMT (Foxa1)
Muscle cell differentiation (Tbx1, Lbx2, MyoD1)
Neuron differentiation (Fzd2, En1, Bhlhe22, Vax1)
Artery development (Tbx1, Hes1)

Figure 4



GO term	Enriched genes	FDR
Killing by host of symbiont cells	ROMO1, DEFA5, CFHR1	0.03
Regulation of neuroinflammatory response	CD200R1, MMP3, PTGS2	0.03

GO term	Enriched promoters	FDR
Leukocyte tethering or rolling	SELL,SELE,SELP	0.03

GO term	Enriched genes	FDR
Positive regulation of multicellular organismal process	GHSR, MESP1, FOXD1, MDK, ZNF703, **CXCR4**, ADRB1, FOXA2	0.05