



# Report

 汇报人: Lilian

2022/3/24



## 文献来源

*The Annals of Applied Statistics*

2013, Vol. 7, No. 1, 523–542

DOI: 10.1214/12-AOAS597

© Institute of Mathematical Statistics, 2013

# JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES<sup>1</sup>

BY ERIC F. LOCK, KATHERINE A. HOADLEY, J. S. MARRON  
AND ANDREW B. NOBEL

*University of North Carolina at Chapel Hill*

联合和个体变化解释 (JIVE) 用于多种数据类型的综合分析



## Abstract & Introduction

- ✓ 许多科学研究领域现在分析高维数据，其中针对给定的一组实验对象测量大量变量。这些数据越来越多地包括一组通用对象的多个高维数据集。

Field	Object	Data types
Computational biology	Tissue samples	Gene expression, microRNA, genotype, protein abundance/activity
Chemometrics	Chemicals	Mass spectra, NMR spectra, atomic composition
Atmospheric sciences	Locations	Temperature, humidity, particle concentrations over time
Internet traffic	Websites	Word frequencies, visitor demographics, linked pages



## Abstract & Introduction

- ✓ 本文的动机是对生物数据的特殊应用。在生物医学研究中，许多技术现在通常收集有关生物体或组织样本的各种信息。来自多种平台和技术的可用生物数据量正在迅速扩大。2011 年核酸研究在线数据库集合列出了 1330 个公开可用的数据库，这些数据库测量了分子和细胞生物学的各个方面 ([Galberin 和 Cochrane, 2011 年](#))。大型在线数据库，例如 ArrayExpress ([Parkinson et al., 2009](#)) 和 UCSC Genome-browser ([Rhead et al., 2010](#)) 通常包含从一组通用样本中收集的多种数据类型。[人类基因组计划](#) ([Sporns、Tononi 和 Kotter, 2005 年](#)) 和癌症基因组图谱 ([TCGA 研究网络, 2008 年](#)) 等大型项目专注于对多种数据类型的综合分析。

## Abstract & Introduction

癌症基因组图谱 (TCGA) 包括来自相同癌性肿瘤样本的几种不同基因组技术的数据。在本文中，我们介绍了联合和个体变异解释 (JIVE)，分解由三个项组成：**捕获跨数据类型的联合变化的低秩近似、针对每个数据类型的结构化变化的低秩近似以及残余噪声。**

- ✓ JIVE 量化数据类型之间的联合变化量，降低数据的维数，并为关节和个体结构的视觉探索提供了新的方向。
- ✓ 所提出的方法包括了主成分分析的扩展，并且对比典型的相关分析和偏最小二乘法等流行的两块方法具有明显的优势。对多形性胶质母细胞瘤肿瘤样本的基因表达和 miRNA 数据的 JIVE 分析揭示了基因-miRNA 的关联，并提供了对肿瘤类型的更好表征。

## Data

特别关注一组 234 个多形性胶质母细胞瘤 (GBM) 肿瘤样本。GBM 是一种常见且非常致命的恶性脑肿瘤，了解肿瘤样本之间的系统差异可能制定更有针对性的治疗方案。

在本文中，我们专注于 miRNA 和基因表达数据的综合分析。miRNA 主要作为基因表达的转录后调节剂发挥作用，被认为是负调节剂，降低基因表达水平。**最近的研究表明，miRNA 可能部分负责众所周知的肿瘤激活基因（癌基因）和肿瘤抑制基因的表达。**

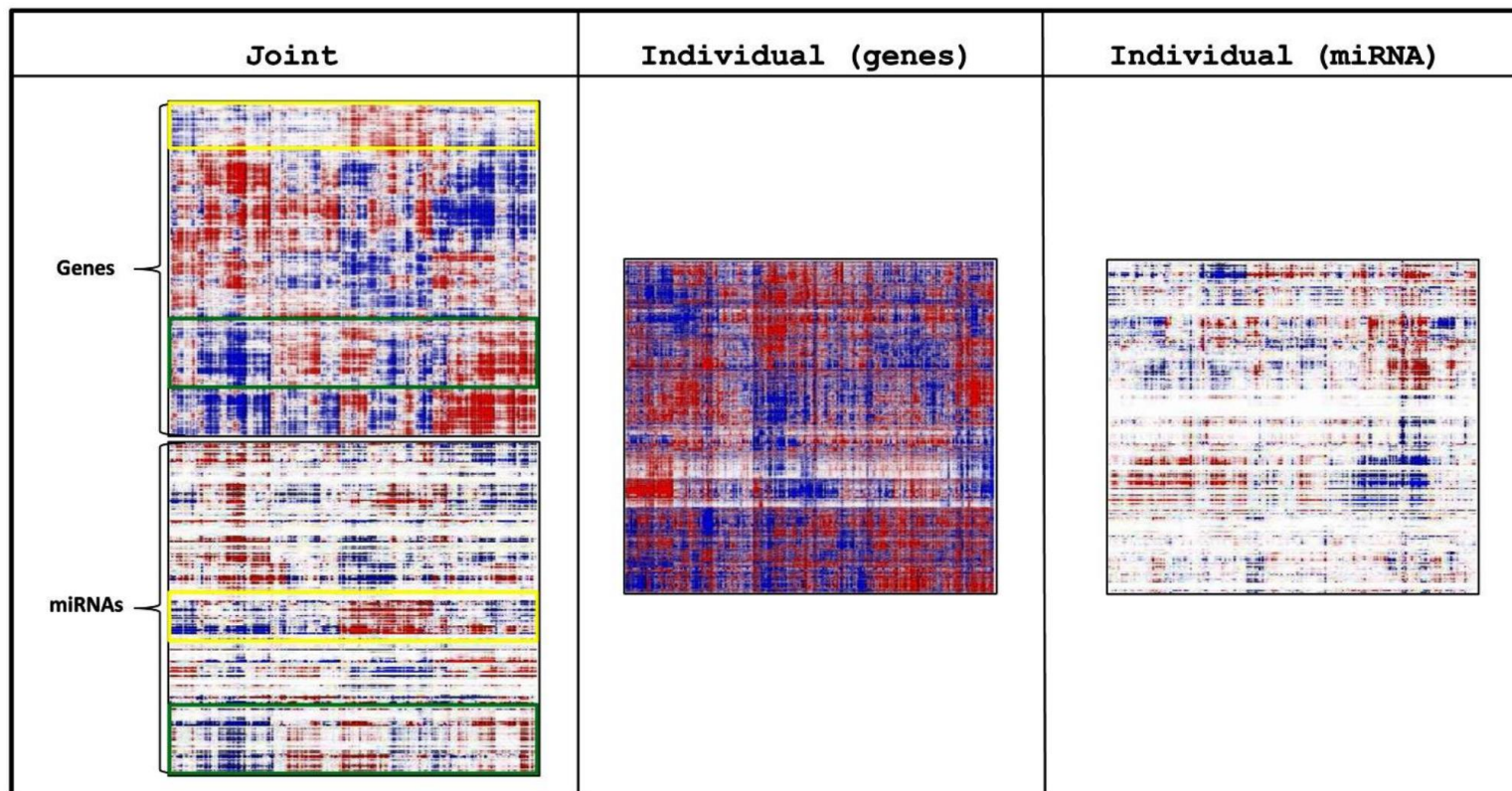
对于每个肿瘤样本，有 534 个 miRNA 和 23,293 个基因（信使 RNA）的强度测量值。这些数据可从 TCGA 公开获得。

## Proposed method.

鉴于基因表达和 miRNA 之间的生物学关系，可以合理地预期两组测量中的共享模式。我们将这种共享模式称为联合结构 (*joint structure*) 。

为了分离联合效应和个体效应，我们引入了一种称为联合和个体变化解释 (JIVE) 的方法。这种探索性方法将数据集分解为三个项的总和：**捕获数据类型之间的联合结构的低秩近似、捕获每个数据类型的单独结构的低秩近似和残余噪声**。对单个结构的分析提供了一种方法来识别存在于一种数据类型中但不存在于其他数据类型中的潜在有用信息。考虑单个结构还可以更准确地估计数据类型之间的共同点。如第 4.2 节所示，JIVE 可以识别现有方法未发现的关节结构。无论数据集的维度是否超过样本大小，都可以使用它。





JIVE 估计 GBM 数据中的关节结构和个体结构。蓝色对应负值，红色对应正值。列在联合结构中以相同的顺序给出，共享相似模式的基因和 miRNA 子集以绿色和黄色突出显示。



## Model and Estimation

例如，在我们对 GBM 数据的应用中， $X$  的行1包含基因表达测量值（维度  $p_1 = 23,293$ ）和  $X$  的行2包含同一组 234 个组织样本（ $n = 234$ ）的 miRNA 测量值（维度  $p_2 = 534$ ）。

$$\begin{aligned}
 X &= \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} : p \times n, & X_i^{\text{scaled}} &= \frac{X_i}{\|X_i\|} & X^{\text{scaled}} &= \begin{bmatrix} X_1^{\text{scaled}} \\ \vdots \\ X_k^{\text{scaled}} \end{bmatrix}. \\
 & & \xrightarrow{\quad \quad \quad} & & & \\
 & & \|A\|^2 = \sum_{i,j} a_{ij}^2 & & & \\
 & & & & \downarrow & \\
 & & & & \|X_i^{\text{scaled}}\| &= 1
 \end{aligned}$$

## Model and Estimation

Let  $X_1, X_2, \dots, X_k$  be matrices as in Section 2.1, scaled appropriately. Joint structure is represented by a single  $p \times n$  matrix of rank  $r < \text{rank}(X)$ . Individual structure for each  $X_i$  is represented by a  $p_i \times n$  matrix of rank  $r_i < \text{rank}(X_i)$ .

More formally, let  $A_i$  be the matrix representing the individual structure of  $X_i$ , and let  $J_i$  be the submatrix of the joint structure matrix that is associated with  $X_i$ . Then, the unified JIVE model is

$$\begin{aligned} X_1 &= J_1 + A_1 + \varepsilon_1 \\ &\vdots \\ X_k &= J_k + A_k + \varepsilon_k, \end{aligned} \tag{2.2}$$

where  $\varepsilon_i$  are  $p_i \times n$  error matrices of independent entries with  $\mathbb{E}(\varepsilon_i) = 0_{p_i \times n}$ . Let

$$J = \begin{bmatrix} J_1 \\ \vdots \\ J_k \end{bmatrix}, \quad \rightarrow \quad \left[ \begin{array}{l} \text{秩约束} \\ JA_i^T = 0_{p \times p_i} \end{array} \right]$$

## Model and Estimation

固定秩 $r$ 、 $r_1$ 、...、 $r_k$ 进行模型估计。通过最小化平方误差之和来估计联合和个体结构。在考虑联合和个体结构后，令 $R$ 为残差的 $p \times n$ 矩阵：

$$R = \begin{bmatrix} R_1 \\ \vdots \\ R_k \end{bmatrix} = \begin{bmatrix} X_1 - J_1 - A_1 \\ \vdots \\ X_k - J_k - A_k \end{bmatrix}.$$

We estimate the matrices  $J$  and  $A_1, \dots, A_k$  by minimizing  $\|R\|^2$  under the given ranks. This is accomplished by iteratively estimating joint and individual structures:

- Given  $J$ , find  $A_1, \dots, A_k$  to minimize  $\|R\|$ .
- Given  $A_1, \dots, A_k$ , find  $J$  to minimize  $\|R\|$ .
- Repeat until convergence.

## Rank Selection.

为了测试联合结构，我们置换了每种数据类型中的列（跨所有行），这保持了每种数据类型的多元分布，同时有效地消除了数据类型之间的关联。为了测试单个结构，我们对数据类型的每一行中的列进行置换，这样可以保持每个变量的分布，同时有效地消除变量间的关联。

- (1) To estimate  $r$ , with  $n\_perm$  permutations and  $\alpha \in (0, 1)$  (by default,  $n\_perm = 100$  and  $\alpha = 0.05$ ):
  - (a) Let  $\lambda_j$  be the  $j$ 'th singular value of  $X = [X'_1 \dots X'_k]'$ ,  $i = 1, \dots, \text{rank}(X)$ .
  - (b) Permute the columns within each  $X_i$ , and calculate the singular values of the resulting concatenated matrix. Repeat  $n\_perm$  times.
  - (c) Let  $\lambda_i^{\text{perm}}$  be the  $100(1 - \alpha)$  percentile among the  $j$ 'th singular values after permutation.
  - (d) Choose  $r$  to be the largest integer such that  $\forall j \leq r, \lambda_j > \lambda_j^{\text{perm}}$ .

ingular values after permuting the columns within each data type. The rank  $r$  is chosen such that for  $j \leq r$ , the  $j$ 'th singular value in the original data is greater than the  $100(1 - \alpha)$  percentile of the distribution of the  $j$ 'th singular value under permutation. The parameter  $\alpha$  is a significance threshold that is applied to each of the singular values.

ied by com-  
with the sin-



## Model and Estimation

- Initialize  $X^{\text{Joint}} = X = [X'_1 \dots X'_k]'$
- Loop:
  - Estimate  $J = [J'_1 \dots J'_k]'$  by a rank  $r$  SVD of  $X^{\text{Joint}}$  ( $J = U\Lambda V'$ )
  - For  $i = 1, \dots, k$ :
    - \* Set  $X_i^{\text{Individual}} = X_i - J_i$
    - \* Estimate  $A_i$  by a rank  $r_i$  SVD of  $X_i^{\text{Individual}}(I - VV')$
    - \* Set  $X_i^{\text{Joint}} = X_i - A_i$
  - Set  $X^{\text{Joint}} = [X_1^{\text{Joint}'} \dots X_k^{\text{Joint}'}]'$



## Relation to PCA

The JIVE model can be factorized as in *Principal Component Analysis* (PCA). For a row-centered  $p \times n$  matrix  $X$ , the first  $r$  principal components yield the rank  $r$  approximation

$$X \approx US,$$

where  $S(r \times n)$  contains the sample scores and  $U(p \times r)$  contains the variable loadings for the first  $r$  components.

As in PCA, the rank  $r$  joint structure matrix  $J$  in the JIVE model can be written as  $US$ , where  $U$  is a  $p \times r$  loading matrix and  $S$  is an  $r \times n$  score matrix. Let

$$U = \begin{bmatrix} U_1 \\ \vdots \\ U_k \end{bmatrix}$$

where  $U_i$  gives the loadings of the joint structure corresponding to the rows of  $X_i$ . The rank  $r_i$  individual structure matrix  $A_i$  for  $X_i$  can be written as  $W_i S_i$ , where  $W_i$  is a  $p_i \times r_i$  loading matrix and  $S_i$  is an  $r_i \times n$  score matrix. Then, the low-rank decomposition of  $X_i$  into joint and individual structure is given by  $X_i \approx U_i S + W_i S_i$ . This gives the factorized model

$$\begin{aligned} X_1 &= U_1 S + W_1 S_1 + R_1 \\ &\vdots \\ X_k &= U_k S + W_k S_k + R_k. \end{aligned} \tag{3.1}$$

Joint structure is represented by the common score matrix  $S$ . These scores summarize patterns in the samples that explain variability across multiple data types. The loading matrices  $U_i$  indicate how these joint scores are expressed in the rows (variables) of data type  $i$ . The score matrices  $S_i$  summarize sample patterns individual to data type  $i$ , with variable loadings  $W_i$ .

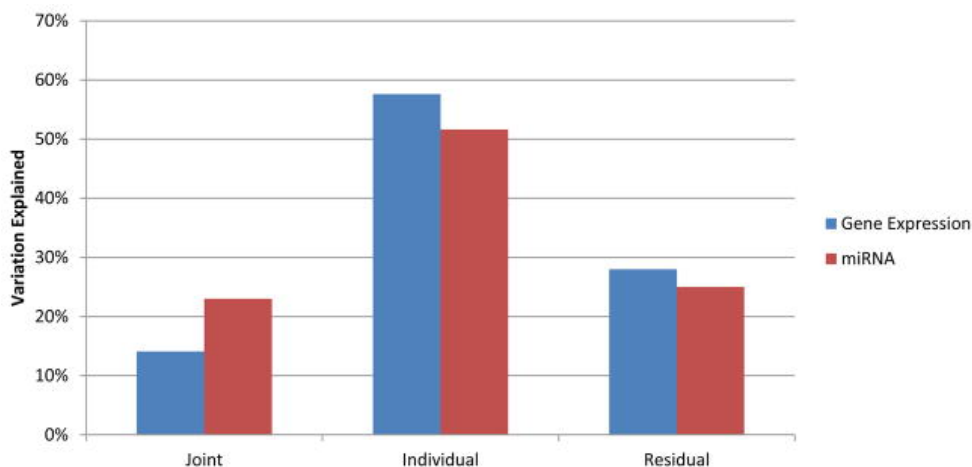


## GBM 数据

由于 GBM 样本的基因表达和 miRNA 数据在维度和可变性上不同，因此按第 2.1 节对它们进行了缩放。排列测试（见附录 A）用于确定估计的关节和个体结构的等级。确定的测试（使用  $\alpha = 0.01$  和 1000 个排列）

- 5级关节结构
- 排名 33 结构个体对基因表达。
- 排名第 13 的结构个体到 miRNA。

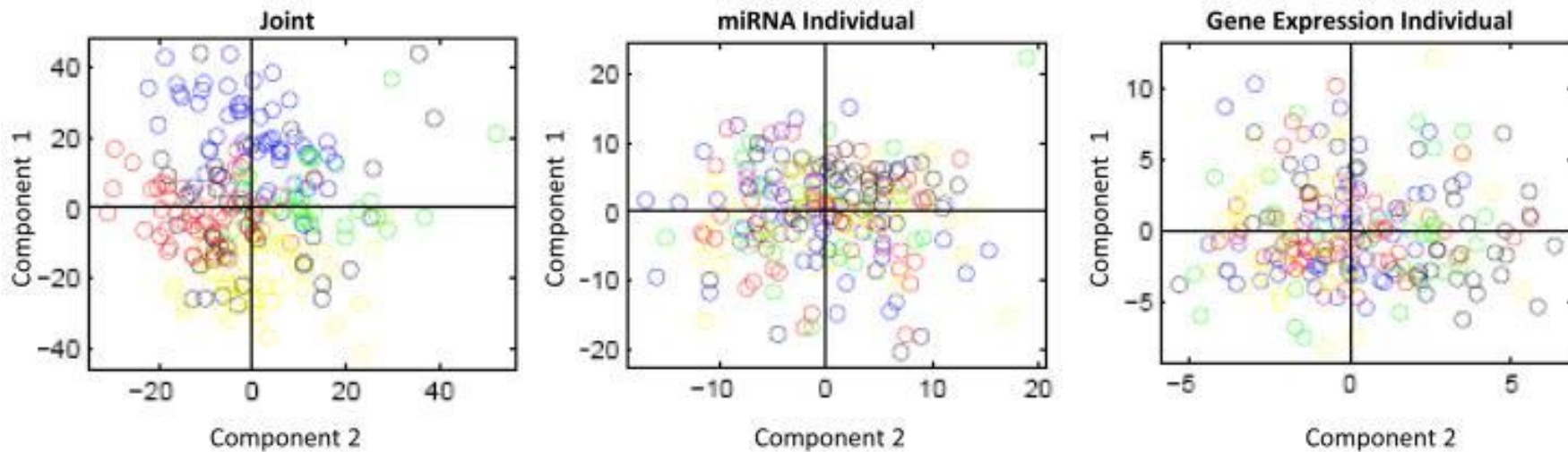
与基因表达相比，联合结构对 miRNA 的变异更多（分别为 23% 和 14%），并且基因表达数据中存在大量与 miRNA 无关的结构化变异（58%）。这与当前的生物学理解是一致的，因为 miRNA 只是影响基因表达的几个因素之一。



由 miRNA 和基因表达数据的估计关节结构、个体结构和残余噪声解释的变异百分比（平方和）



## GBM 数据



前两个联合成分、前两个个体 miRNA 成分和前两个个体基因表达成分的样本分数散点图。样品按亚型着色：间充质（黄色）、原神经（蓝色）、神经（绿色）和经典（红色）。在初始亚型分析后对黑色样本进行了分析，并被认为是未分类的。

## Variable Sparsity

在 JIVE 分解中使用惩罚项来诱导变量稀疏性。如果关节和单个结构的一些可  
变载荷恰好为 0，则实现稀疏性。对于权重  $\lambda$  和  $\lambda_i$ ，最小化惩罚平方和

$$\|R\|^2 + \lambda \text{Pen}(U) + \sum_i \lambda_i \text{Pen}(W_i),$$

➤ Pen 是一种惩罚，旨在诱导加载向量中的稀疏性。

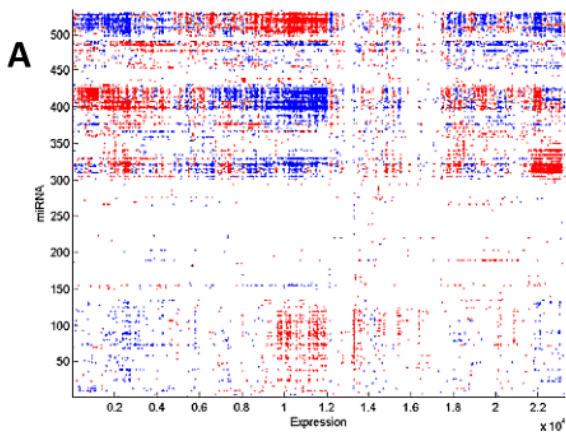
在这种惩罚下，具有较小或不显着贡献的变量的负载趋于缩小到 0。其他稀疏  
诱导惩罚（例如硬阈值）可以代替 L1 惩罚。

Estimation with sparsity is accomplished by an iterative procedure analogous to  
that used for the nonsparse case:

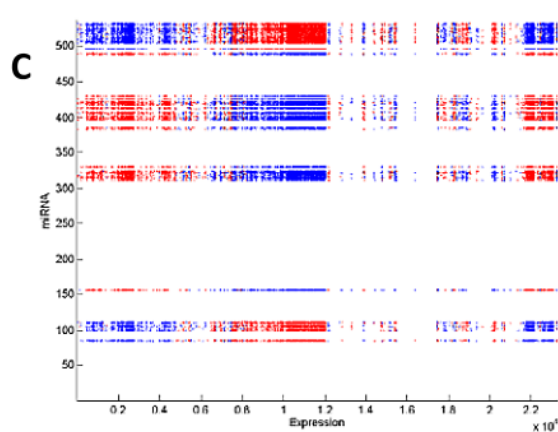
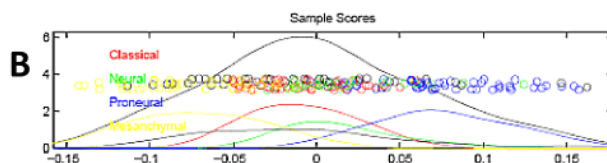
- Given  $J$ , find  $A_i$  to minimize  $\|R_i\|^2 + \lambda_i \text{Pen}(W_i)$  for each  $i = 1, \dots, k$ .
- Given  $A_1, \dots, A_k$ , find  $J$  to minimize  $\|R\|^2 + \lambda \text{Pen}(U)$ .
- Repeat until convergence.



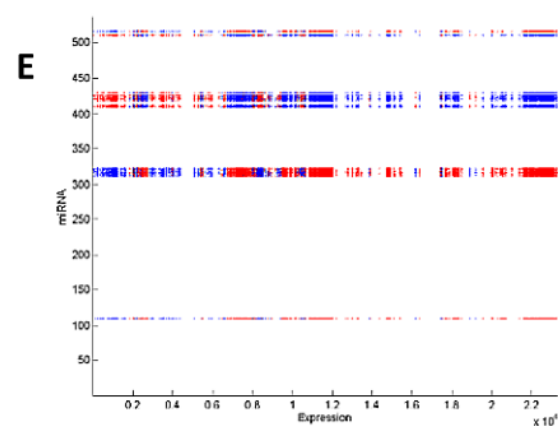
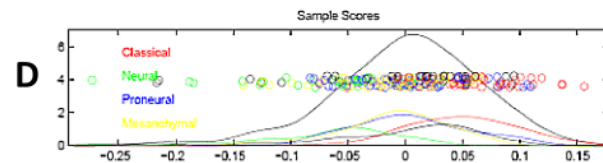
## Correlations



## Component 1

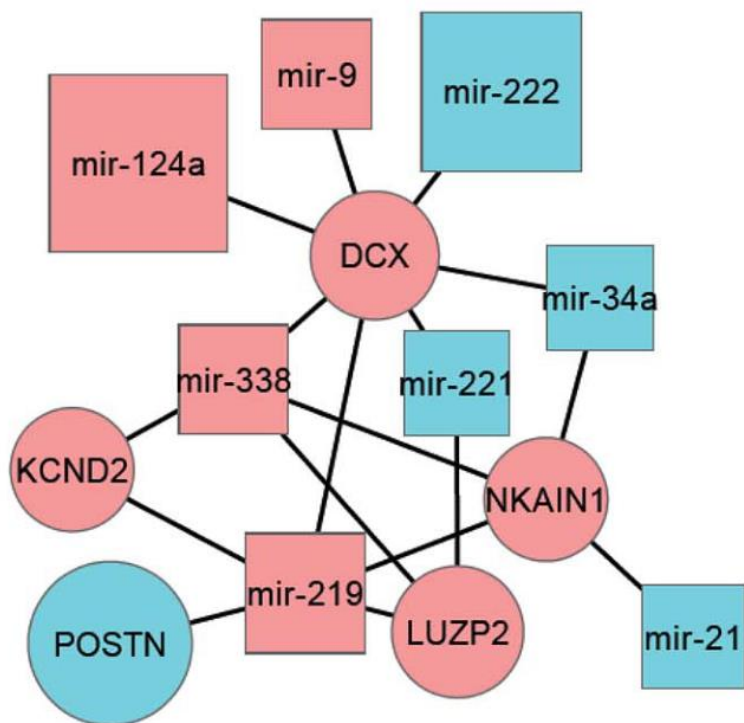


## Component 2

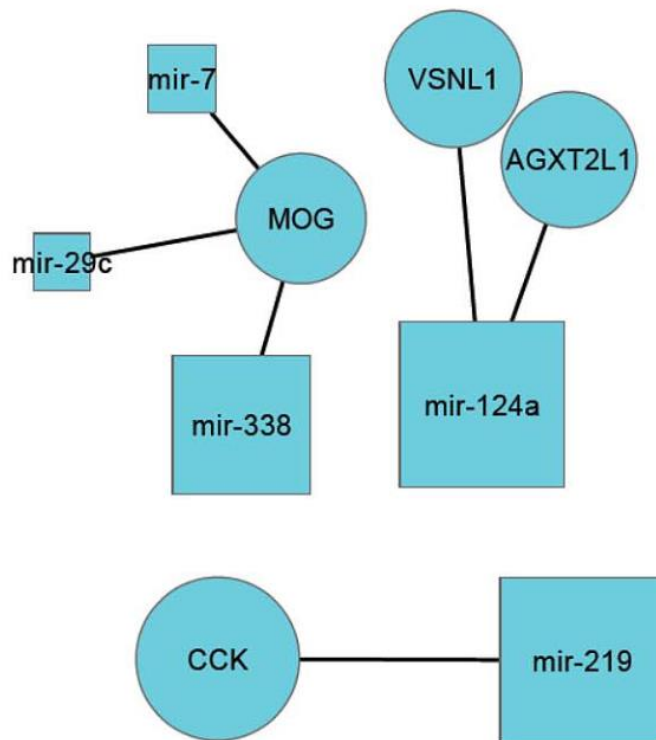




Component 1



Component 2



进一步检查了对前两个关节成分贡献最大的单个基因和 miRNA。POSTN 基因在第一个关节成分的基因中负载最大，它编码蛋白质 Periostin。

Periostin 的过度表达经常在癌性肿瘤细胞中被报道，并且被怀疑促进细胞运动（癌细胞快速和自发迁移的能力）。

## Conclusions

JIVE 既发现了多种数据类型的协调活动，也发现了特定数据类型独有的那些特征。我们演示了如何考虑联合结构可以更好地估计单个结构，反之亦然。我们将 JIVE 应用于 GBM 肿瘤样本的基因和 miRNA 数据，可以更好地表征肿瘤类型，并更好地理解给定数据类型之间的生物相互作用。

虽然本文侧重于垂直整合的生物医学数据，但 JIVE 模型和算法非常通用，可能在其他情况下有用。类似的方法可以应用于水平整合的数据，其中不同的样本集（例如生病和健康的患者）可用于相同的数据类型。在金融领域，JIVE 有可能改进当前解释不同市场之间和内部变化的模型








# nature

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [articles](#) > [article](#)

Article | [Published: 09 March 2022](#)

## The evolution, evolvability and engineering of gene regulatory DNA

[Eeshit Dhaval Vaishnav](#) , [Carl G. de Boer](#) , [Jennifer Molinet](#), [Moran Yassour](#), [Lin Fan](#), [Xian Adiconis](#), [Dawn A. Thompson](#), [Joshua Z. Levin](#), [Francisco A. Cubillos](#) & [Aviv Regev](#) 

[Nature](#) **603**, 455–463 (2022) | [Cite this article](#)

**26k** Accesses | **1** Citations | **508** Altmetric | [Metrics](#)

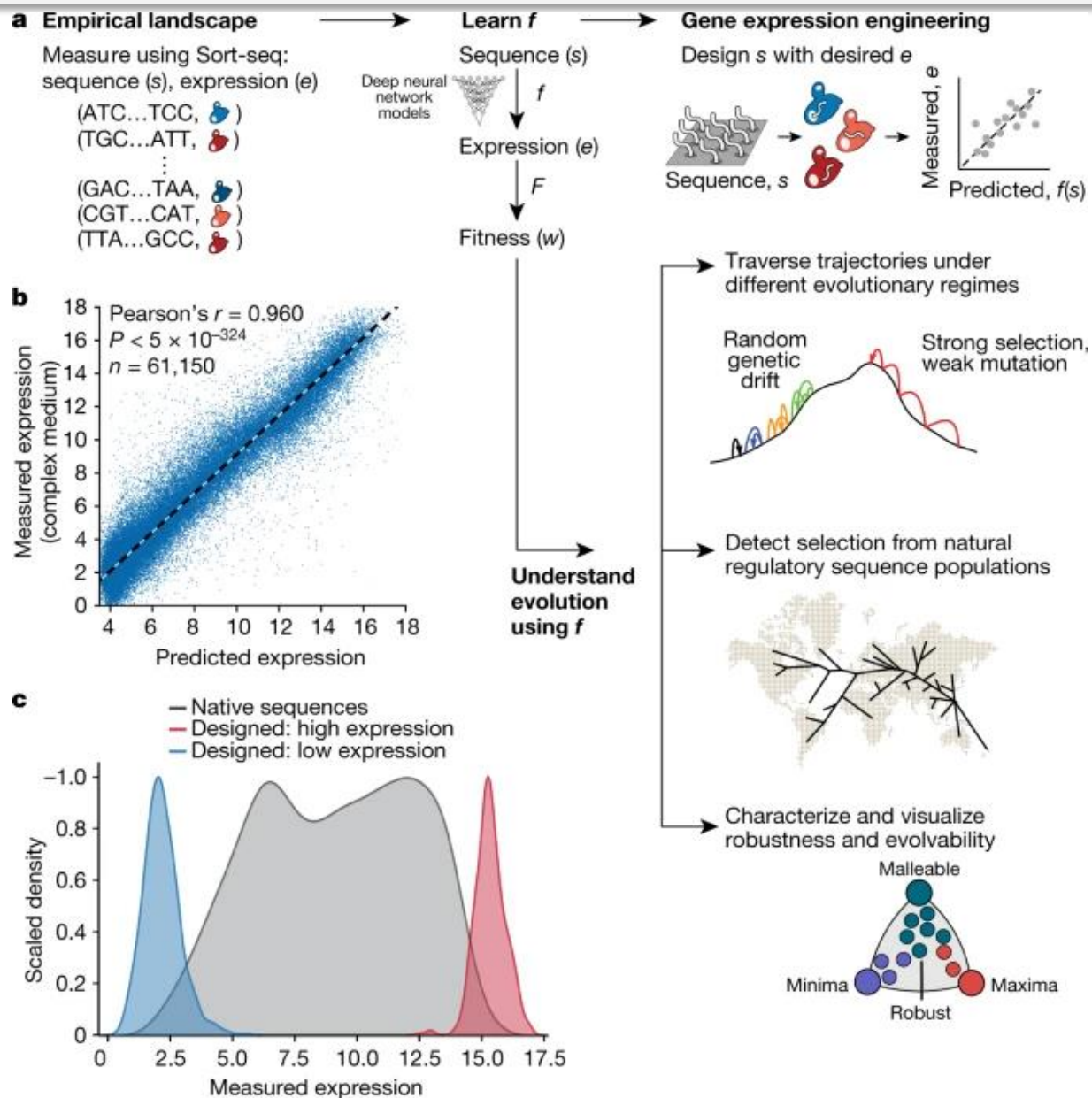
基因调控 DNA 的进化、可进化性和工程

作者首先合成了超过3000万个携带长度为80个碱基的随机酵母启动子序列，然后嵌入同一含有YFP荧光基因的报告系统中

## 本文概述

卷积神经学习模型不仅能够实现对未知酵母启动子序列的转录强度的精准预测，更重要的是，其在预测任务上的极佳表现直接反映了与转录强度相关的启动子本身的内在序列特征的存在性。

作者成功地实现了对这些序列依据进化潜力和进化模式特征的相似性进行降维和可视化，并发现了三个独立而互补的能够共同解释这些进化特征差异性的“原型”（archetype）。





## nature communications

Article | [Open Access](#) | Published: 03 December 2019

# Estimating heritability and genetic correlations from large health datasets in the absence of genetic data

Gengjie Jia, Yu Li, Hanxin Zhang, Ishanu Chattopadhyay, Anders Boeck Jensen, David R. Blair, Lea Davis, Peter N. Robinson, Torsten Dahlén, Søren Brunak, Mikael Benson, Gustaf Edgren, Nancy J. Cox, Xin Gao & Andrey Rzhetsky 

*Nature Communications* **10**, Article number: 5508 (2019) | [Cite this article](#)

从大规模电子病历中估算疾病遗传参数

## 本文概述

本研究的主要目的是在不引入新的遗传数据的情况下，利用现有的电子病历和遗传参数，通过机器学习的方法为500多种的疾病来估计其遗传率和遗传相关性。

作者通过大型健康医疗数据集构建了两类疾病指标：

- 1) 疾病患病率曲线，
- 2) 疾病嵌入。前者记录了疾病的动态趋势，而后者寻求用并发症来表征该疾病。这两类指标相辅相成，能够用来：

- 1) 估算数百种疾病用相似的算法**word2vec**实现。利用该算法的目的是为疾病建立实值向量表示，以实现在给定当前疾病的情况下预测它的并发疾病，或给定并发疾病来预测当前疾病情况。和数千种疾病对的遗传参数，

- 2) 系统地分析遗传率与疾病发病年龄的关系，

- 3) 并将疾病患病率曲线形状的差异与疾病之间的遗传和环境相关性联系起来。

此外，该研究团队还发布了可搜索的网络共享资源，其中包括了500多种疾病的性别和国家特异的流行曲线。

最后得到的预测准确度可以与传统的基于大量遗传数据研究中得到的结果相媲美。

## Article

# Tumour DDR1 promotes collagen fibre alignment to instigate immune exclusion

<https://doi.org/10.1038/s41586-021-04057-2>

Received: 18 October 2019

Accepted: 22 September 2021

Published online: 03 November 2021

 Check for updates

Xiujie Sun<sup>1,11</sup>, Bogang Wu<sup>1,11</sup>, Huai-Chin Chiang<sup>1,11</sup>, Hui Deng<sup>2</sup>, Xiaowen Zhang<sup>1</sup>, Wei Xiong<sup>2</sup>, Junquan Liu<sup>2</sup>, Aaron M. Rozeboom<sup>3</sup>, Brent T. Harris<sup>3</sup>, Eline Blommaert<sup>4</sup>, Antonio Gomez<sup>5</sup>, Roderic Espin Garcia<sup>4</sup>, Yufan Zhou<sup>6</sup>, Payal Mitra<sup>7</sup>, Madeleine Prevost<sup>1</sup>, Deyi Zhang<sup>8</sup>, Debarati Banik<sup>1</sup>, Claudine Isaacs<sup>3</sup>, Deborah Berry<sup>3</sup>, Catherine Lai<sup>3</sup>, Krysta Chaldekas<sup>3</sup>, Patricia S. Latham<sup>9</sup>, Christine A. Brantner<sup>10</sup>, Anastas Popratiloff<sup>10</sup>, Victor X. Jin<sup>6</sup>, Ningyan Zhang<sup>2</sup>, Yanfen Hu<sup>7</sup>, Miguel Angel Pujana<sup>4</sup>✉, Tyler J. Curiel<sup>8</sup>✉, Zhiqiang An<sup>2</sup>✉ & Rong Li<sup>1</sup>✉

## Nature: DDR1——三阴性乳腺癌治疗新靶点

研究发现利用三阴性乳腺癌小鼠模型发现了DDR1 (Discoidin Domain Receptor 1) 调控肿瘤发生发展的机制，并为三阴性乳腺癌的治疗提供了新思路。

- DDR1促进具有免疫活性的宿主的乳腺肿瘤生长
- DDR1抑制抗肿瘤免疫细胞的浸润
- DDR1依赖性的细胞外基质重构抑制抗肿瘤免疫浸润
- DDR1作为肿瘤免疫治疗的治疗靶点

血管外渗、  
通过肿瘤诱导的趋化性  
细胞外基质ECM