# Saver

1. 背景

本文使用的Drop-seq[1]是一种广泛应用的单细胞RNA测序技术，基于液滴微流控，简单来说就是将细胞悬浮液与带有barcode的beads泵至微流控芯片中，两者汇聚后被油滴包裹，形成单分散的微滴droplet，随后将细胞裂解，释放mRNA并与beads上的barcode杂交，后续逆转录，以及PCR扩增，然后就可以进行测序和分析。单细胞水平的RNA测序相比bulk测序提供了更精细的分辨率揭示不同细胞之间的差异，探索细胞群体内的细胞异质性。

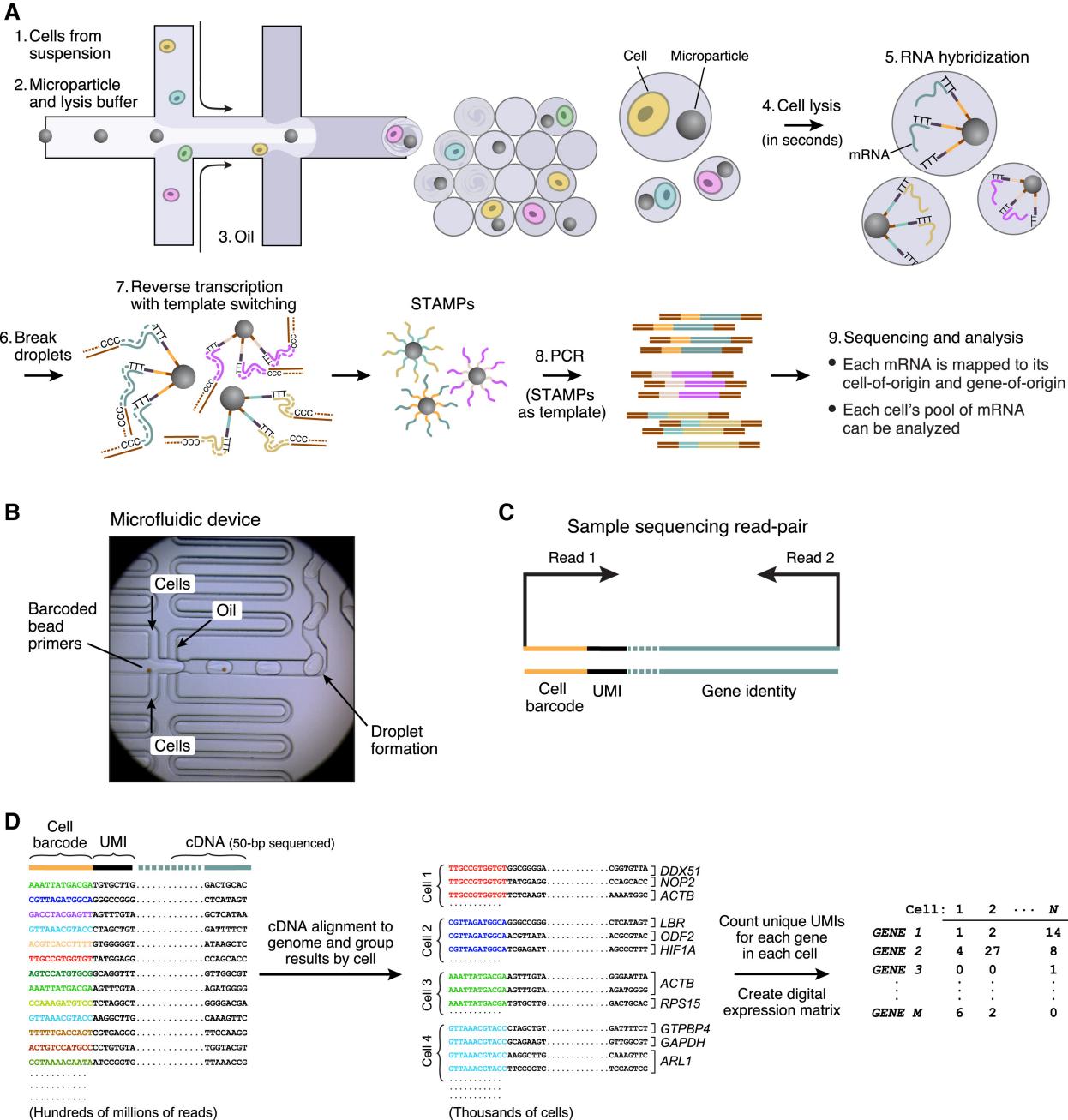


图1 Drop-seq测序流程[1]

但由于scRNA-seq技术中的低RNA输入量以及PCR扩增等步骤，导致数据存在较高的噪音和技术变异，在有限的测序深度下，scRNA-seq数据是高度稀疏，但又高维度的，特别是本身表达量低的基因，可能在大部分细胞中都无法检测到，本身表达量高的基因可能又受限于测序深度上限，导致读数饱和。

基于此，scRNA-seq中，reads数的分布通常会被建模为负二项分布或泊松分布的变体，因为方差与均值之间并非简单的线性关系，通常来说表达量更高的基因，方差更收缩且大于均值，而表达量更低的基因，由于受噪音影响更大，方差常常离散度更高-方差异质性。[2]

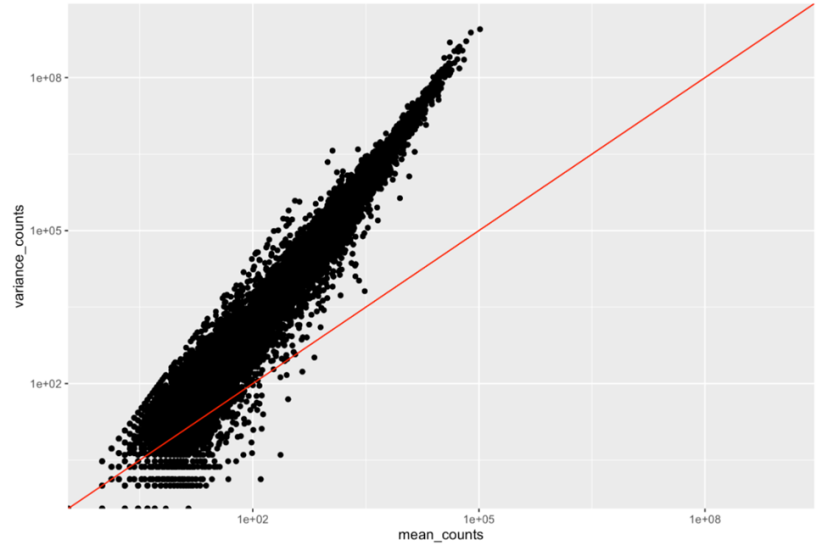
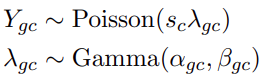


图2 RNA-seq数据的reads分布特点[2]

1. 方法和结果
2. SAVER的主要思想：

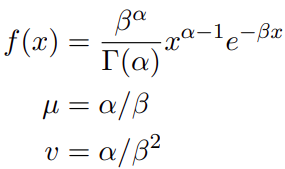
将scRNA-seq数据建模为泊松-伽马分布，假设每个基因的计数数据服从泊松分布，考虑到数据的离散性，引入伽马分布作为泊松分布的参数的先验分布，但作者并没有直接指定Gamma先验，而是用部分高质量基因的表达作为真实信号，从数据中学习参数的先验分布，并估计先验参数（gamma分布重新参数化后的μ和v）。一旦先验参数被估计出来，SAVER就可以得到所有基因表达的预测值，对预测值和观察值进行加权平均后，就可以输出真实表达的后验分布，最后不仅用后验分布的平均值作为SAVER恢复的表达值，且量化了估计的不确定性。

1. 模型假设

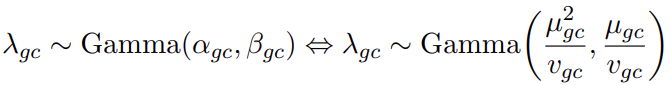


Ygc是细胞c中基因g的UMI数，对Ygc用gamma-poisson建模。Sc是细胞特异的归一化常数，λgc是真实的表达量。α和β是gamma分布中的形状参数和尺度参数。

Gamma分布表达式如下：



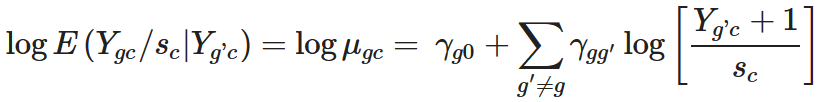
μ和v分别表示Gamma分布的随机变量X~Gamma(α, β)的均值和方差。用μ和v对gamma分布重新参数化：



接下来的目标就是估计先验均值μ和先验方差v。

1. 参数估计
2. 先验均值μ的估计

均值估计：μgc表示对基因g的预测值，g’表示来自相同细胞中的其它基因。将g'的log归一化count数作为泊松广义线性回归模型的predictors，来预测基因g的表达值。



在Lasso回归中，惩罚项λ被添加到似然函数中控制predictors的数量，一个较大的λ会使得模型有大量的零值系数，也就是只有少数基因作为predictors，对g基因的表达（μgc）有贡献。用选定的高表达基因，利用glmnet回归，选择五折交叉验证中误差最小的惩罚项λ对应的模型，利用该模型预测每个基因g的先验表达均值μ^gc.

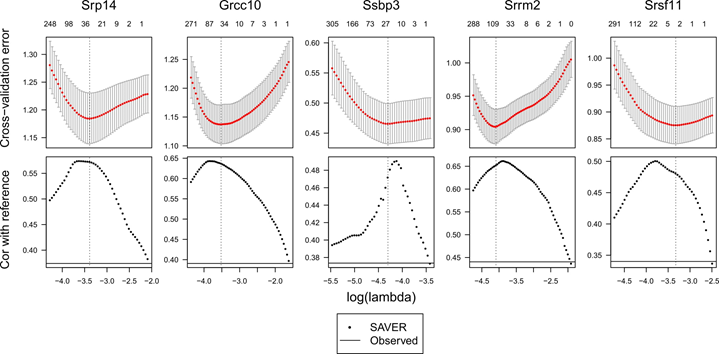
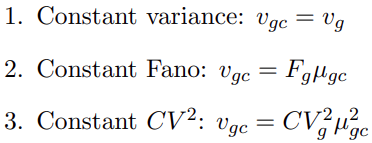


图3 五折交叉验证的损失

1. 先验方差v的估计

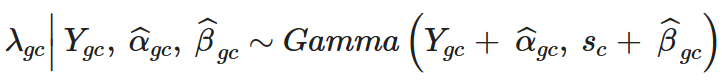
通过假设一个由离散参数ϕg表示的细胞间的恒定噪声模型来估计先验方差。该模型根据均值-方差关系的不同假设，有以下三种：分别表示方差为常数，方差和均值呈线性关系，以及方差和均值呈二次关系。



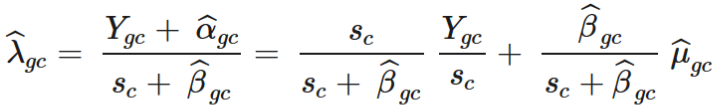
为了保证在给定µgc和vgc情况下，观察表达量Ygc出现的概率最大，分别求上述三种噪声模型的密度函数，然后最大化似然函数log值，找到最大值对应的v^gc。

1. 表达量的预测

得到μ^gc，v^gc和对应的噪声模型Φg，对gamma分布重新参数化，后验分布变为：



然后计算该分布的平均值作为SAVER对该基因表达量的估计值：



Ygc是观测值，μgc是估计值（预测值）

通过β^gc项控制，该项表示均值/方差，当ϕg^小，也就是低噪音，β^gc大时，后验均值主要由权重更大的预测值μˆgc决定，反之亦然。

1. 模型评估
2. 相对表达量

作者用FISH数据集和Drop-seq实验数据集重叠的基因来评估SAVER的效果。

首先是计算了Gini系数，一种表征数据分布和异质性的指标，来评估SAVER对表达量分布的恢复效果，图4可以看到，以FISH结果作为真实信号的情况下，相比原始Drop-seq数据，MAGIC插补数据，scImpute插补数据，SAVER插补后计算的Gini系数，与FISH计算得到的Gini系数相关性更高且均方根误差更小，即SAVER能够有效的恢复基因表达的分布特征。

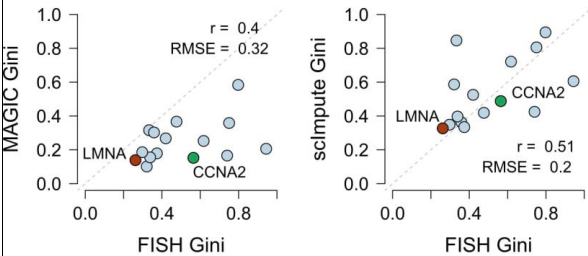
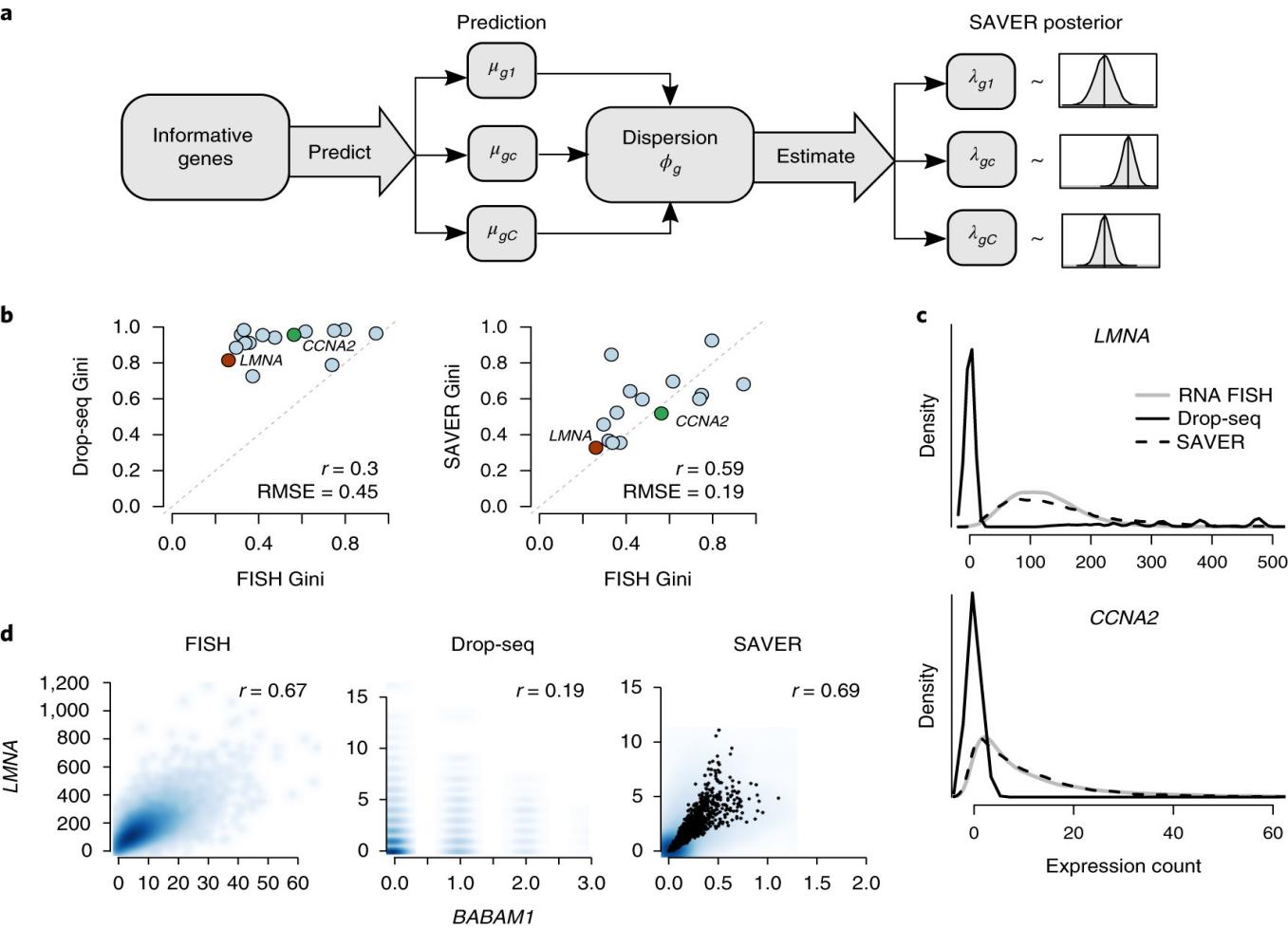


图4 不同方法得到的数据之间Gini系数的相关性

除此之外，SAVER还能更好的恢复真实的基因间的相关性，如图5，SAVER计算的相关性值与FISH的相关性值更一致，且减少了噪声，相比MAGIC引入大量虚假的相关性，SAVER更可靠地估计真实相关性为0的基因的关联。（图5）

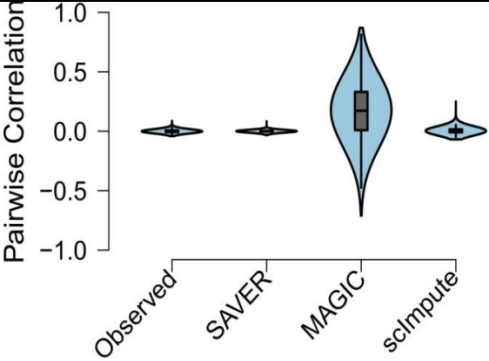
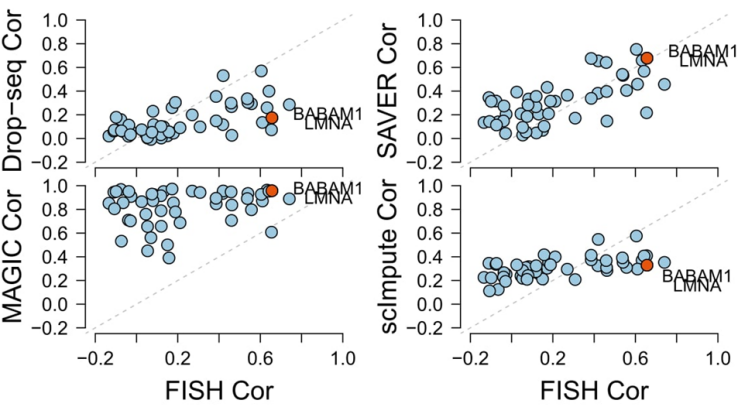


图5 不同方法得到的基因间的Pearson相关性（左）及真实零关联基因的相关性（右）的比较

1. 绝对表达量及下游分析

鉴于很难确定每个细胞中mRNA的实际数量，作者对数据集进行了下采样实验，生成基准数据集。首先选择具有高表达的基因和细胞作为reference，将这些表达水平视为真实表达。然后，以低效率方式模拟捕获和测序过程，获取观察数据集，接着运行SAVER，MAGIC和scImpute，以及其他插补的算法获得恢复数据集。

为了评估每种方法的性能，作者计算了reference和观察数据以及恢复数据集之间的细胞间和基因间Pearson相关系数。如图6，SAVER改善了基因间和细胞间的相关性。除此之外，作者还研究了SAVER对下游分析的影响，包括差异表达和细胞聚类分析。SAVER在下采样数据集中检测到了最多的差异表达基因，同时控制了FDR。

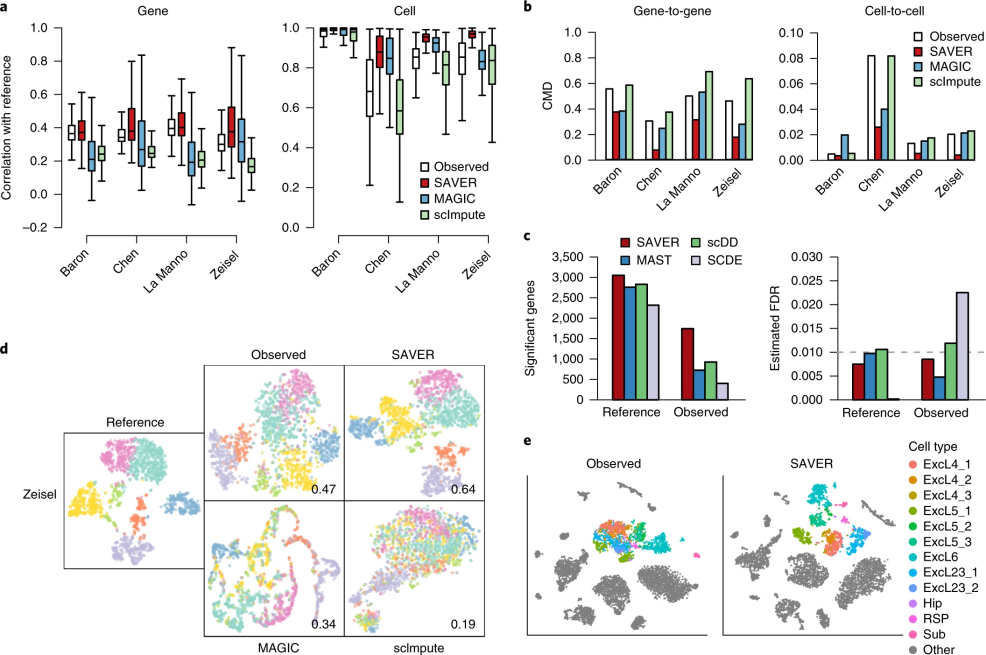


图6 下采样后表达值的恢复及对下游分析的影响

1. Macosko E , Basu A , Satija R , et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets[J]. Cell, 2015, 161(5):1202-1214.
2. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.