营养与健康所等建立MAnorm2计算模型

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/11940.html

近期，*Genome Research*在线发表了中国科学院上海营养与健康研究所中科院计算生物学重点实验室（马普伙伴计算生物学研究所）研究员邵振课题组的方法学论文——*MAnorm2 for quantitatively comparing groups of ChIP-seq samples*，报道了其开发的新一代MAnorm2计算模型。该模型能够对多样本ChIP/ATAC-seq数据按照特定标签分组，进行统计建模和组间定量比较，可靠地在样本组层面鉴定组间显著差异的ChIP/ATAC-seq信号。

染色质免疫共沉淀测序（ChIP-seq）实验被广泛用于刻画转录因子结合和组蛋白修饰的全基因组分布。比较来自不同细胞类型的ChIP-seq样本是刻画细胞分化及病变过程中动态转录和表观调控的关键基础。2012年，邵振与中科院分子植物科学卓越创新中心研究员张一婧等合作，在*Genome Biology*上发表了用于两个ChIP-seq样本之间进行一对一定量比较的MAnorm模型。近年来，随着实验技术的发展和测序成本的不断降低，在ChIP-seq样本组（而非单个样本）之间进行比较分析，已成为越来越常见的研究需求。一方面，科研人员会产生同一实验的多个生物学重复以提高实验结果的可信度；另一方面，通过将来自不同个体的样本根据特定标签（如年龄、性别、患病与否、疾病亚型等）分组进行比较，研究人员能够控制个体差异造成的影响，更可靠地识别与该标签关联的差异结合位点。然而，由于ChIP-seq实验固有的高复杂度和高噪声水平，以及不同比较场景特有的技术困难，现阶段对多样本ChIP-seq数据进行分组定量比较，仍是一个较大的计算方法学挑战。

在ChIP-seq数据标准化这一步，MAnorm2沿用了MAnorm的核心假设，通过重构其信号强度变换体系，新发展了以参照样本为基准的多样本并行ChIP-seq信号标准化流程。进一步，针对多样本分组比较的需求，MAnorm2搭建了一个理论上适应任意树状分组结构的层级化多样本标准化策略。在完成标准化后，MAnorm2针对每个基因组区域上观察到的ChIP-seq信号组间差异进行统计检验。在通常组内样本数较少的局限下（2～3个重复本），为更准确地衡量每一个基因组区域上的组内样本间ChIP-seq信号变化水平（within-group variability），MAnorm2设计出一个经验贝叶斯框架，利用拟合均值-方差曲线以给单个区域的组内变化水平赋予一个先验分布，并进一步通过平衡先验和后验观测以更准确地估计ChIP-seq信号的组内变化水平，从而提高对组间差异ChIP-seq信号的灵敏度（图1）。

与已有的其他经验贝叶斯方法相比，MAnorm2的最大优势在于考虑了不同样本组的组内ChIP-seq信号变化水平可能存在系统性差别。这一情形在正常人和癌症患者之间的比较中常出现：由于肿

瘤组织或血液样本本身的异质性及癌症亚类型和不同患病阶段的多样性，癌症样本组的组内信号变化水平高于正常样本组。为解决该问题，MAnorm2通过在建模过程中引入一个方差比率因子，把不同样本组的全局组内信号变化水平修正至一致，使用修正后的方差进行均值-方差曲线的拟合和参数估计（图2）。研究人员系统比较了MAnorm2与现有的其他ChIP-seq差异分析工具，发现MAnorm2展现出更优越的使用性能，尤其是当进行比较的样本组拥有不同组内变化水平时，如癌症和正常样本作比较。此外，该模型的应用场景和统计模型具有良好的可扩展性。研究人员展示了MAnorm2在ATAC-seq数据差异分析上同样适用，将其统计模型扩展至可同时比较的任意多个样本组，发现其使用效果优于传统的ANOVA方法。

营养与健康所博士后涂世奇为论文第一作者，邵振为论文通讯作者。张一婧、美国西南医学中心教授徐剑、波士顿大学教授David J. Waxman对该研究工作的提出和完善做出了重要贡献。研究工作获得国家自然科学基金委、科技部、中科院等的支持。

论文链接

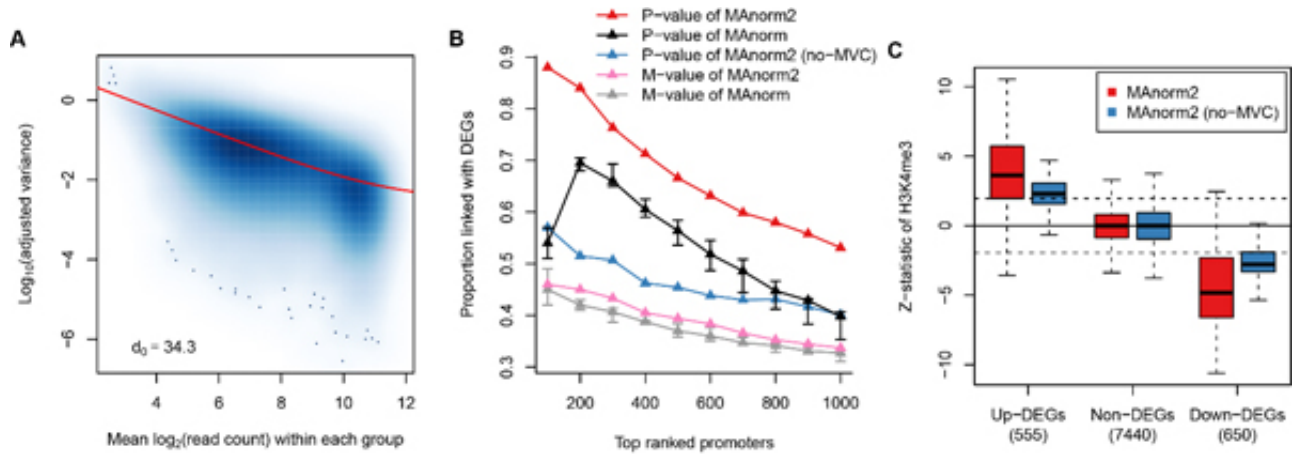

图1.（A）在不同基因组区域间拟合均值-方差曲线（mean-variance curve；MVC）。（B）根据不同的统计指标对基因启动子按照差异H3K4me3 ChIP-seq信号的可能性进行排序，并计算其中差异表达基因（differentially expressed genes；DEGs）启动子所占的比例。（C）检查不同类型的基因启动子上差异H3K4me3的统计显著性。虚线对应P值为0.05

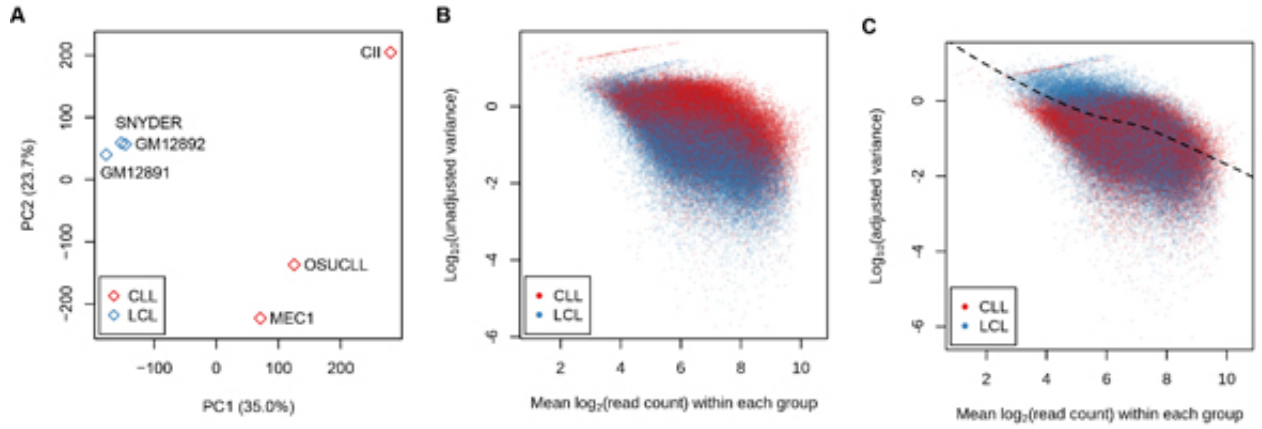

图2.（A）对来自不同的人的H3K27ac ChIP-seq样本进行主成分分析。这里LCL（lymphoblastoid

cell line）组包含三个源于正常人的B细胞的细胞系；CLL（chronic lymphocytic leukemia）组包含三个源于慢性淋巴细胞白血病患者的B细胞的细胞系。（B）关于来自不同组的均值和未修正的方差的散点图。（C）关于均值和修正后的方差的散点图，以及由此进行下一步统计建模

研究团队单位：上海营养与健康研究所