QOE：新型生物信息数据挖掘工具

冀开元、马文丽、郑文岭

广东省广州市南方医科大学基因工程研究所 510515

摘要：随着生物芯片技术的发展，生物信息数据呈指数增长。如何有效地分析挖掘数据成为生物医药目前急需解决的问题。Qlucore Omics Explorer(QOE)是一款新的生物信息分析软件，可用于快速分析基因表达、基因芯片、实时 PCR 以及 DNA 甲基化等多种生物学数据。本文对 Qlucore Omics Explorer (QOE)的基本功能和特点做了介绍，包括软件背景，数据输入输出类型，软件界面，软件应用等，最后分析了 QOE 的应用前景。

关键字：QOE，生物信息，基因表达

QOE, a new bioinformatics data mining software

Kaiyuan Ji, Wenli Ma, Wenling Zheng

Institute of Genetic Engineering, South Medical University, Guangzhou, 510515, China

Abstract: With the advancement of gene microarray, there is an exponential increase in bioinformatics data. To explore the bioinformatics data and get meaningful analysis become urgent in biomedical studies now. Qlucore Omics Explorer is a new software for analyzing gene expression, microarray, real-time PCR, and DNA methylation as well as other kinds of bioinformatics data. We review the basic function and characteristics of Qlucore Omics Explorer, including the background, the import and export of data, the window of software, application of software, as well as the future perspectives of the applications of the Qlucore Omics Explorer.

Keywords: Qlucore Omics Explorer, bioinformatics, gene expression

1. QOE 背景

基因芯片自问世以来，以其强大的力量迅速席卷生物和医学领域[1]。基因芯片使得研究者可以从全基因组水平对基因表达谱、药物代谢、疾病发生发展过程进行快速的定量分析[2]。随着现代分子生物学的发展，以及人类基因组计划的完成，生命科学的研究已经进入后基因组时代。近年来芯片技术的革新和优化，使现在芯片数据急剧增长，芯片质量大幅度提高。生物学家面对的不再是零散的、少量的、简单的数据，而是公共数据库中数以万兆计的、各种各样的复杂生物数据，GEO(Gene Expression Omnibus)公共数据库中数据集的数据也在逐年增长[3]（图 1）。各个研究所和实验室产生出了大量的实验数据，但是这些数据的信息往往没有被完全挖掘出来。因此，如何有效分析这些生物数据成为当前的生物学瓶颈，这就迫使人们寻求一种有效的方法和工具去管理筛选这些数据，并且对它们进行统计、聚类和进一步利用。海量的生物学数据中必然蕴含着重要的生物学规律，这些规律将是解释生命之谜的关键。

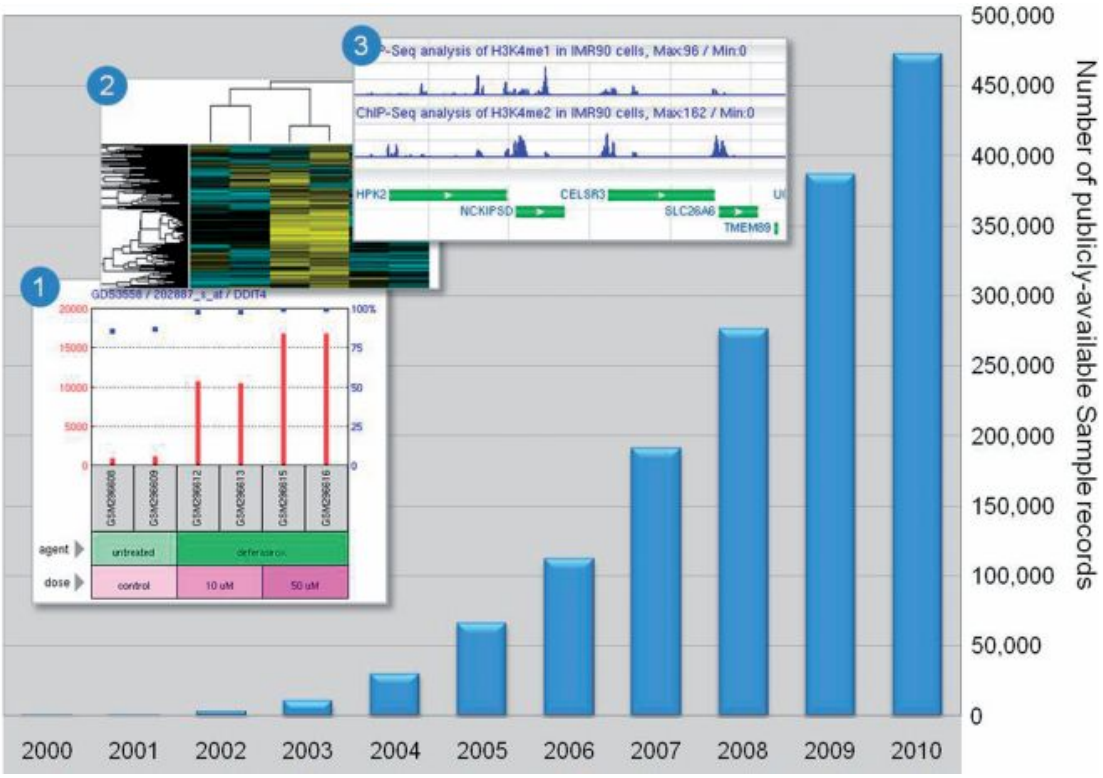

图 1．GEO 数据库中样本数量的增长情况

Figure1．The growth of sample records in GEO

由于生物信息数据庞大复杂，数据成千上万，变量也不仅仅是几百个几千个，而是上万个，并且涉及多个学科的交叉[4]。通常需要研究者精通生物学、计算机软件学和统计学相关专业，实际上生物学家往往不懂计算机专业知识，而且计算机专业的学者又没有生物学的思维，即使两者合作，也会有技术上的代沟。因此，最佳的解决方法是给生物学家一些方便快捷的生物信息学分析工具，通过简化的统计学方法来挖掘生物信息数据集。

传统的生物信息学分析软件常常不容易学习和掌握，多数生物信息分析软件也仅能提供简单的数据统计信息并且操作麻烦、分析速度慢，在面对大规模数据集时有心无力。Qlucore Omics Explorer 是瑞典隆德大学的合作研究项目，由数学和医学遗传学系的研究人员研究开发而成，用于处理表达谱芯片产生的大量的高

维数据。此软件把数据经统计计算转换为 3D 可视化图，再经过主成分分析和聚类降维[5, 6]，从多个角度识别出数据中隐藏的结构和模式。QOE 软件巧妙的编程和设计，使得用户可在普通电脑上交互实时地探索和分析高维数据集。QOE 把大量数据统计和分析功能进行了系统融合，形成了类似视窗浏览器的简单界面，使生物信息数据分析过程更为简便、直观和形象，从而满足了生物学工作者的生物信息学分析需求。

2. QOE 功能

QOE 作为一款生物信息学数据分析软件，可以快速、直观地显示分析后的数据图示，同时应用各种生物统计学原理，对数据进行深入和有效地分析。QOE 只需要一个普通的电脑就可以实时处理庞大的数据集（超过 100 万个条目）。操作 QOE 不需要对数学或统计有深入的了解，也不需要具备一台强大的超级电脑，只需要一台普通电脑，就能够轻松探索高维数据并迅速得到相关结果。QOE 具有很强的易用性和快速性，可以直接在电脑屏幕上实时对数据进行可视化处理。

QOE 能直接引用标准化的 Agilent 公司和 Affymetrix 公司的芯片数据。也可以直接从公共数据库 GEO 中下载所需数据并导入到软件中。使用者可以直接导入自己的数据集也可以导入网上下载的其它数据集，其中 GEO 数据库的数据集可以直接在 QOE 中输入相应 GSE 或 GDS 编号来导入，并且选择性自动加载相应注释文件。该软件能同时打开多个数据集，共同来比较分析。QOE 能将数据进行实时的 2D 和 3D 的可视化演示，所有的图都可以完全互动。下图所示为将软件自带示例数据——急性白血病亚型的数据集[7]导入后的分析界面（图 2）。导入的数据集可在工作窗口即时生成可视化的图，包括主成分分析（principal

component analysis）图，散点（scatter）图，线（line）图，热（heat）图，箱（box）图等，这个分析过程极大地方便了研究者进一步寻找数据间的规律和关系，并且所有的图都会随着操作即时更新。

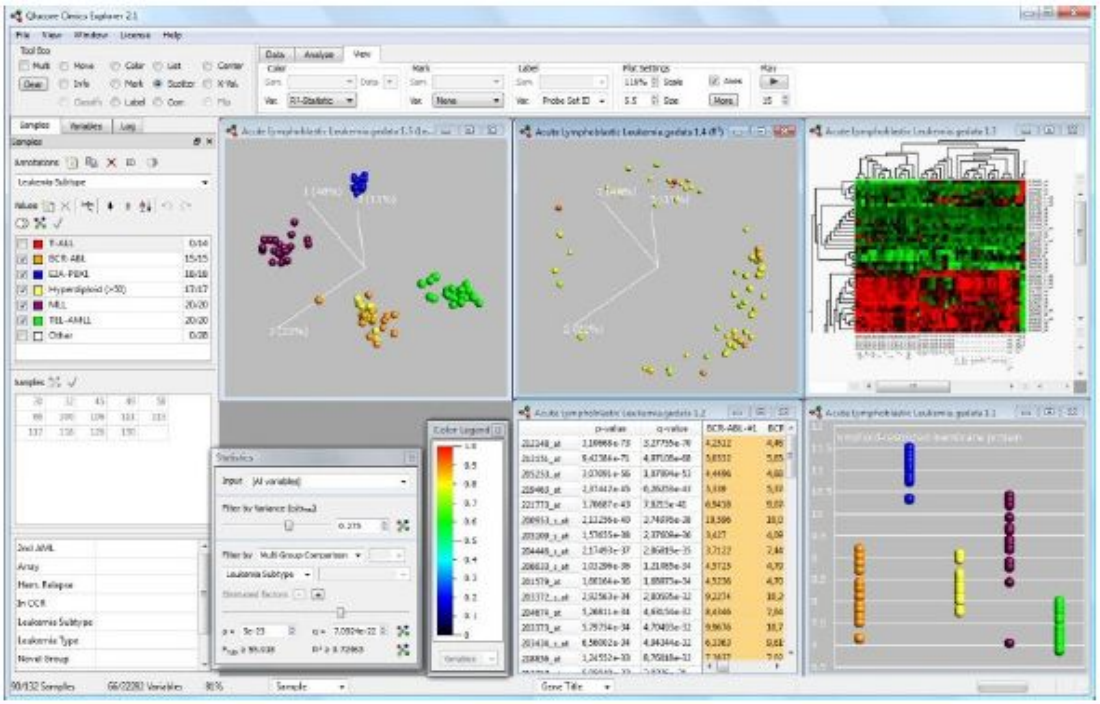

图 2：Acute Lymphoblastic Leukemia data set 的 QOE 分析结果.

Figure2：Analysis results of Acute Lymphoblastic Leukemia data set on QOE

数据导入后，QOE 会把数据集的原始值进行 mean=0 以及 var=1 的标准化处理，并且进一步把每个变量计算出相应的标准误 σ 。使用者可以通过统计窗口中的 σ / σ_{max} 来筛选掉样本之间差异小的变量，PCA 图和热图会随着操作实时变化，使用者可以通过图的变化选择其它统计学方法进行下一步的分析。

在整个分析过程中，QOE 根据不同目的进行不同的统计方法和措施，及时计算出相应的 p 值和 q 值，通过点击鼠标来改变过滤器的参数，从而探索数据筛选出所需目的基因或者发现新的结构和关系。主要的统计方法有：两组比较(two group comparison)，多组比较(multi group comparion)，线性回归（linear

regression ），二次回归（quadratic regression），秩回归（rank regression）。而且还有 fold change 功能去进一步供我们选择应用。改变过滤变量参数的设置，有助于研究人员找到不同生物学统计意义条件下的生物学方面数据的改变，从而有利于展开进一步的实验研究。

在分析的过程中，内置的 GSEA（Gene Set Enrichment Analysis）和 GO（Gene Ontology）插件，可帮助使用者分析基因功能和通路信息。可以对重要的变量或样本进行着色和标记，而且可以根据自我需求导入或导出不同的注释信息列表，如 GI、gene symbol、gene ID、gene title 等。产生的图也可以随时的导出，并可以导出 PCA 的视频。操作的进度可以随时保存，保证了分析的连贯性。QOE 的官方网站为 http://www.qlucore.com/，在此网站可以观看操作视频以及相关文件，并且提供针对微软 window 操作系统 32 位和 64 位的两种版本的 QOE，目前的版本名称为 Qlucore Omics Explorer 3.0。

3. QOE 支持的文件

QOE 支持直接导入的标准化的文件包括：Affymetrix 公司和 WT 的阵列、安捷伦基因阵列、安捷伦 microRNA 阵列。

对于 Illumina 的数据，建议是通过 GenomeStudio 或 BeadStudio 软件将数据标准化，然后使用向导将数据导入到 Qlucore。其它的芯片仪器产生的数据或来自其它类型的数据，大部分都可以导入到 Qlucore 中。支持的全部格式如下：

文件类型	文件后缀

Affymetrix 公司的的 CEL 文件	.cel
Affymetrix 的探针组文件	.chp
安捷伦文本文件	.txt
安捷伦基因查看文件	.txt
简单的数据文件	.txt
Qlucore 数据文件	.gedata
GEO 数据集	.soft 和.soft.gz
GEO 系列矩阵	.txt 和.txt.gz
小型文本文件	.txt;.csv
生物芯片软件文件（	.base

4. QOE 的界面

打开QOE，双击电脑桌面上的QOE图标(快捷方式)或者从电脑的Start Menu（开始菜单）中的 Program Menu（程序菜单）打开 QOE。QOE 的 Main Window（主窗口）将出现，如图 3。

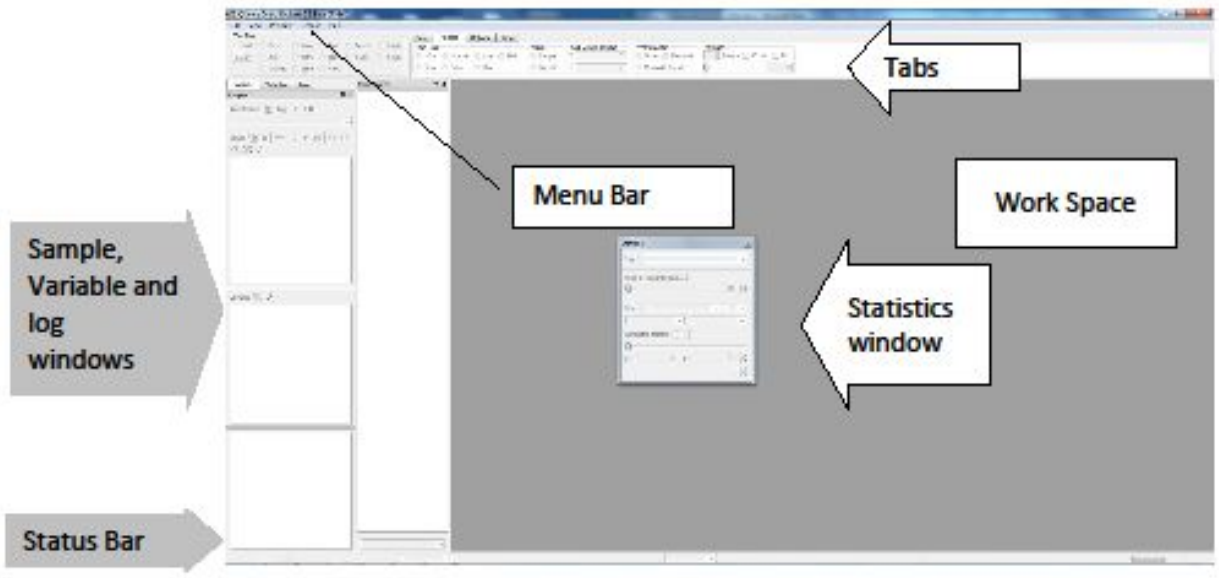

图 3：QOE 软件的操作界面。

Figure3: The window of QOE.

在屏幕中间，会显示在 Plot Windows（图形窗口）的 Work Space（工作空间）。此外，会发现几个 dock windows（停靠窗口）。在默认状态下，Samples（样本）、Variables（变量）和 Log（日志）等三个停靠窗口停靠在主窗口的左侧，而 Statistics（统计）和 Getting Started（入门）等两个窗口则是浮动状态。可以在 Menu Bar（菜单栏）中通过点击 view（视图）> Dock Windows（停靠窗口），选择显示的停靠窗口。菜单栏下可以找到管理，以及在数据集上执行各种操作功能的不同控件。还可以找到能帮助您在 QOE 中选择和管理工作流程的四个不同选项卡（Tabs）：Data（数据），Method（方法），Options（选项）和 View（视图）。在 Statistics dock window（统计停靠窗口）中，可以根据研究数据集的需要选择合适的统计方法。最后，在底部有 Status Bar（状态栏）。在状态栏里会显示例如数据集里样本的总数和各种变量，以及此刻能参与分析的有用信息。

5. QOE 的应用

QOE 既可以用来做实验前的预测和筛选，以确定实验方案。也可以做实验结果的证明，发现未知功能和关系，更适合利用庞大的数据库做生物信息“干实验”。可应用于分析基因表达芯片数据、临床数据、蛋白芯片数据、抗体数据、microRNA 芯片数据、蛋白质芯片数据、实时定量 PCR 数据、DNA 甲基化数据等。只要符合软件要求，任何数据都可以导入到 QOE 中进行分析。

6. 总结和展望

目前由于生物芯片的研究和发展，大量的生物信息比如基因芯片的数据产生出来，而且大量的基因数据也带来了大量的生物信息，这些信息都是些高维数据，基因数据的维数差异给后续分析带来了困难[8]。而 QOE 能通过方差过滤和 PCA 技术相结合的方式解决数据的降维问题。把结果通过 PCA 图的方式呈现出来，经人脑的高效率的分辨挖掘，从而快速有效地分析出有用的信息，避免有用信息被筛选掉，克服了机械处理数据的弊端。

综上所述，QOE 是一个强大的交互式分析和可视化的分析软件。它可用于许多不同类型的数据集进行分析。QOE 支持用户快速、动态地分析和验证各种不同的假说，结合统计检验提供即时的可视化结果。QOE 还可以发现数据中隐藏的结构和大型数据集的隐藏的模型，充分利用各种注释以及各个环节与数据的连接，快速轻松地导出许多不同类型的报告和演示文稿中使用的数据、图像和动画，并进行更深一步研究。QOE 可以使生物学家不依赖于计算机和统计专业的支持，研究自己的数据集。用户界面的设计是直观和易于使用的，可以对数据集

结构进行随心所欲的研究，同时提供了内置功能互动和简单的假设检验。QOE 中的主成分分析图，散点图，热图和数据表也是重要的基本操作，使得高维数据能够低维可视化。QOE 可帮助研究者突破计算机方面的瓶颈，充分利用自己的专业知识完成对数据的挖掘和探索。而且 QOE 还能用于其它高维复杂数据的分析，有着广阔的应用范围。相信在不久的将来，QOE 在生物学、医学、农学等领域的应用会取得很大进展。

References:

[1]. Bellazzi, R., et al., Data analysis and data mining: current issues in biomedical informatics. Methods Inf Med, 2011. 50(6): p. 536-44.

[2]. Schulze, A. and J. Downward, Navigating gene expression using microarrays--a technology review. Nat Cell Biol, 2001. 3(8): p. E190-5.

[3]. Barrett, T., et al., NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res, 2011. 39(Database issue): p. D1005-10.

[4]. Hasman, A., et al., Biomedical informatics--a confluence of disciplines? Methods Inf Med, 2011. 50(6): p. 508-24.

[5]. Ji, H., et al., Differential principal component analysis of ChIP-seq. Proc Natl Acad Sci U S A, 2013. 110(17): p. 6789-94.

[6]. Yeung, K.Y. and W.L. Ruzzo, Principal component analysis for clustering gene expression data. Bioinformatics, 2001. 17(9): p. 763-74.

[7]. Ross, M.E., et al., Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood, 2003. 102(8): p. 2951-9.

[8]. Alter, O., P.O. Brown and D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A, 2000. 97(18): p. 10101-6.

冀开元，男，南方医科大学基因工程研究所在读硕士研究生，主要从事基因表达谱数据分析。