

<<基于Affymetrix芯片的基因表达研究>>

图书基本信息

书名：<<基于Affymetrix芯片的基因表达研究>>

13位ISBN编号：9787030329080

10位ISBN编号：7030329082

出版时间：2012-1

出版时间：科学出版社

作者：〔美〕Hinrich G&ouml;hlmann、Willem Talloen 著，张春秀 译

页数：327

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：<http://www.tushu007.com>

## <<基于Affymetrix芯片的基因 >>

### 内容概要

#### Affymetrix

GeneChip系统是目前应用最广泛的生物芯片平台。

但是由于Affymetrix芯片含有超大量的信息，很多Affymetrix芯片用户趋向于使用默认的分析设置，得到的常常不是最优化的结论。

分子生物学家和生物统计学家根据十余年的基因表达谱实验研究和数据分析的实践经验编写了《基于Affymetrix芯片的基因表达研究》，从理论概念到实验结果，解释了使用Affymetrix芯片进行基因表达研究的全部过程，拆除了分子生物学、生物信息学和生物统计学之间无处不在的语言障碍。

本书权威实用，介绍了Affymetrix芯片的重要技术、统计学易犯的错误和问题，同时涉及其他芯片平台的一般规则和应用。

通过例证和全彩图例，描述了技术和统计方法的概念，为初学者提供详细指导。

本领域的专家则可以了解芯片所涉及的其他学科知识，拓展基因芯片表达谱研究的认识。

# <<基于Affymetrix芯片的基因 >

## 书籍目录

附图目录

表格目录

BioBox目录

StatsBox目录

前言

缩写词和术语

1 生物学问题

1.1 为什么进行基因表达?

1.1.1 生物技术的进展

1.1.2 生物学相关的研究

1.2 研究问题

1.2.1 相关性和实验研究对比

1.3 研究课题的主要类型

1.3.1 两组间比较

1.3.2 多组间比较

1.3.3 不同治疗方式间的比较

1.3.4 多组与对照组的比较

1.3.5 研究主题内的变化

1.3.6 分类和预测样本

2 AffymetriX芯片技术

2.1 探针

2.2 探针组

2.2.1 标准探针组的定义

2.2.2 客户可选择的芯片描述文件(CDF)

2.3 芯片类型

2.3.1 标准表达检测芯片

2.3.2 外显子芯片

2.3.3 基因芯片

2.3.4 叠瓦芯片

2.3.5 用于某项研究的专用芯片

2.4 标准实验室芯片实验流程

2.4.1 体外转录分析

2.4.2 全转录本正义链标记

2.5 AffymetriX芯片的数据质量

2.5.1 分析数据的重复性

2.5.2 分析数据的稳定性

2.5.3 分析的敏感性

3 实验操作

3.1 生物学实验

3.1.1 生物学背景

3.1.1.1 实验目的/假设

3.1.1.2 技术平台

3.1.1.3 mRNA水平的预期改变

3.1.2 样本

3.1.2.1 选择合适的样品/组织

## <<基于Affymetrix芯片的基因 >

- 3.1.2.2 样本的类型
- 3.1.2.3 样本的异质性
- 3.1.2.4 性别
- 3.1.2.5 时间点
- 3.1.2.6 样本切割引起的误差
- 3.1.2.7 动物处理产生的误差
- 3.1.2.8 RNA的质量
- 3.1.2.9 RNA的数量
- 3.1.3 预实验
- 3.1.4 主实验
  - 3.1.4.1 对照实验
  - 3.1.4.2 实验处理
  - 3.1.4.3 分批实验
  - 3.1.4.4 随机化
  - 3.1.4.5 标准化
  - 3.1.4.6 选择对照
  - 3.1.4.7 样品量/重复次数/费用
  - 3.1.4.8 平衡设计
  - 3.1.4.9 对照样本
  - 3.1.4.10 样本混合
  - 3.1.4.11 实验记录
- 3.1.5 实验数据分析验证
- 3.2 芯片实验
  - 3.2.1 外源RNA对照
  - 3.2.2 靶基因合成
  - 3.2.3 批处理影响
  - 3.2.4 全基因组芯片和用于某项研究的专用芯片比较
- 4 数据分析预处理
  - 4.1 数据预处理
    - 4.1.1 探针的信号强度
    - 4.1.2 转换为log<sub>2</sub>的对数
    - 4.1.3 背景校正
    - 4.1.4 归一化
    - 4.1.5 AffymetriX芯片概要
      - 4.1.5.1 完全匹配(PM)和错配(MM)技术
      - 4.1.5.2 只使用PM探针的技术
    - 4.1.6 整体解决方案
    - 4.1.7 信号检测方法
      - 4.1.7.1 芯片分析系统MAS 5.0
      - 4.1.7.2 背景和杂交信号检测(DABG)
      - 4.1.7.3 检出/缺失比值(PANP)
    - 4.1.8 标准化
  - 4.2 质量控制
    - 4.2.1 技术数据
    - 4.2.2 虚拟图像
    - 4.2.3 重复性评价
      - 4.2.3.1 重复性评价方法

## <<基于Affymetrix芯片的基因 >

- 4.2.3.2 实例分析
- 4.2.4 批处理效应
- 4.2.5 批处理效应校正
- 5 数据分析
  - 5.1 为什么我们需要统计学?
    - 5.1.1 需要对数据作出解释
    - 5.1.2 需要一个优秀的实验设计
    - 5.1.3 统计学与生物信息学比较
  - 5.2 高维数据的问题
    - 5.2.1 分析结果的重复性
    - 5.2.2 数据挖掘和验证
  - 5.3 基因过滤
    - 5.3.1 过滤方法
      - 5.3.1.1 信号强度
      - 5.3.1.2 两样品间变异
      - 5.3.1.3 缺失/检出
      - 5.3.1.4 含有效信息的/无有效信息的检出
    - 5.3.2 数据过滤对检验和多重校正的影响
    - 5.3.3 几种过滤方法的比较
  - 5.4 无监督数据分析
    - 5.4.1 进行无监督分析的原因
      - 5.4.1.1 批次影响
      - 5.4.1.2 技术或生物学的偏差
      - 5.4.1.3 表型数据的质量校验
      - 5.4.1.4 共调控基因的识别
    - 5.4.2 聚类
      - 5.4.2.1 距离和联系
      - 5.4.2.2 聚类算法
      - 5.4.2.3 聚类质量校验
    - 5.4.3 多元投影方法
      - 5.4.3.1 多元投影方法类型
      - 5.4.3.2 基因和样本关系图
  - 5.5 检测差异表达
    - 5.5.1 复杂问题的简单解决方法
    - 5.5.2 统计检验
      - 5.5.2.1 倍数变化
      - 5.5.2.2 t-检验类型
      - 5.5.2.3 由t统计到p值
      - 5.5.2.4 方法比较
      - 5.5.2.5 线性模型
    - 5.5.3 多重检验的校正
      - 5.5.3.1 多重检验的问题
      - 5.5.3.2 多重校正步骤
      - 5.5.3.3 方法比较
      - 5.5.3.4 事后比较
    - 5.5.4 统计学意义与生物学相关性
    - 5.5.5 样本数量估计

## <<基于Affymetrix芯片的基因 >

### 5.6 有监督的预测

#### 5.6.1 分类与假设检验

#### 5.6.2 芯片分类的挑战

##### 5.6.2.1 过度拟合

##### 5.6.2.2 偏执方差平衡

##### 5.6.2.3 交叉效验

##### 5.6.2.4 非唯一分类解决方案

#### 5.6.3 位点选择方法

#### 5.6.4 分类方法

##### 5.6.4.1 判别分析

##### 5.6.4.2 最近邻分析法

##### 5.6.4.3 逻辑(Logistic)回归

##### 5.6.4.4 神经网络

##### 5.6.4.5 支持向量机

##### 5.6.4.6 分类树

##### 5.6.4.7 集成方法

##### 5.6.4.8 芯片预测分析(PAM)

##### 5.6.4.9 方法比较

#### 5.6.5 复杂的预测问题

##### 5.6.5.1 多级问题

##### 5.6.5.2 生存预测

#### 5.6.6 样本量

### 5.7 通路分析

#### 5.7.1 通路分析的统计学方法

##### 5.7.1.1 过表达分析

##### 5.7.1.2 功能分类评分

##### 5.7.1.3 基因集分析

##### 5.7.1.4 方法比较

#### 5.7.2 数据库

##### 5.7.2.1 Gene ontology

##### 5.7.2.2 京都基因与基因组百科全书(KEGG)

##### 5.7.2.3 基因芯片通路分析(GenMAPP)

##### 5.7.2.4 腺嘌呤富集元件数据库(ARED)

##### 5.7.2.5 概念图(cMAP)

##### 5.7.2.6 凋亡路径图(BioCarta)

##### 5.7.2.7 染色体位置

### 5.8 其他分析方法

#### 5.8.1 基因网络分析

#### 5.8.2 元分析

#### 5.8.3 染色体位置

### 6 分析结果表示

#### 6.1 数据可视化

##### 6.1.1 热图

##### 6.1.2 强度图

##### 6.1.3 基因表图

##### 6.1.4 维恩图(Venn图)

##### 6.1.5 散点图

## <<基于Affymetrix芯片的基因 >

- 6.1.5.1 火山图(Volcano plot)
- 6.1.5.2 MA图
- 6.1.5.3 高维数据的散点图
- 6.1.6 柱状图
- 6.1.7 盒图
- 6.1.8 小提琴图表
- 6.1.9 密度图
- 6.1.10 树状图
- 6.1.11 基因表达通路
- 6.1.12 出版用图表
- 6.2 生物学解释
  - 6.2.1 重要数据库
    - 6.2.1.1 Entrez Gene
    - 6.2.1.2 Affymetrix网站(NetAffx)
    - 6.2.1.3 OMIM
  - 6.2.2 文献挖掘
  - 6.2.3 数据整合
    - 6.2.3.1 多种分子筛选数据
    - 6.2.3.2 系统生物学
  - 6.2.4 实时定量聚合酶反应(RTqPCR)验证
- 6.3 数据发表
  - 6.3.1 ArrayExpress
  - 6.3.2 基因表达文库(GEO)
- 6.4 可重复性研究
- 7 药物研发
  - 7.1 早期标志物的需求
  - 7.2 关键路径计划
  - 7.3 药物发现
    - 7.3.1 正常组织和病变组织的不同
    - 7.3.2 疾病亚型的发现
    - 7.3.3 分子靶标的识别
    - 7.3.4 分子特征谱
    - 7.3.5 疾病模型特征
    - 7.3.6 化合物分析
    - 7.3.7 剂量效应处理
  - 7.4 药物开发
    - 7.4.1 生物标志物
    - 7.4.2 响应显著性
    - 7.4.3 毒理基因组学
  - 7.5 临床实验
    - 7.5.1 功能指标
    - 7.5.2 结果预测的意义
- 8 使用R和Bioconductor
  - 8.1 R和Bioconductor
  - 8.2 R和Sweave(R语言的一种函数)
  - 8.3 R和Eclipse(一种代码)
  - 8.4 自动芯片分析

## &lt;&lt;基于Affymetrix芯片的基因 &gt;

- 8.4.1 装载文件包
- 8.4.2 基因过滤
- 8.4.3 无监督探索
- 8.4.4 差异表达检验
- 8.4.5 有监督分类
- 8.5 其他芯片分析软件
- 9 未来前景
- 9.1 同时分析不同数据类型
- 9.2 未来的芯片
- 9.3 新一代(二代)测序:芯片的终结?
- 参考文献
- 索引
- 附图目录
- 2.1 标准AffymetriX芯片图
- 2.2 GC含量对信号强度的影响
- 2.3 同一探针集中的探针之间信号强度的差别
- 2.4 使用客户选择的CDF时,探针集大小引起的差异
- 2.5 外显子芯片和3 端芯片探针覆盖范围的比较
- 2.6 外显子芯片的转录本注释
- 3.1 性别特异基因Xist(X染色体失活特异转录本)
- 3.2 样本切割产生误差示例
- 3.3 甲状腺素在小鼠纹状体的表达
- 3.4 小鼠结肠样本切割引起的误差
- 3.5 降解与非降解RNA对比
- 3.6 RNA的降解图显示3 偏差
- 3.7 不同批次芯片的批间效果
- 4.1 芯片扫描图像的一角
- 4.2 对数转换的分配效应
- 4.3 芯片数据中的两种噪音成分
- 4.4 归一化对强度依赖变异的影响
- 4.5 归一化对MA图的影响
- 4.6 MAS 5.0背景计算
- 4.7 由affyPLM产生的虚拟图像
- 4.8 两重复关联评估重复性
- 4.9 中心定位前后的成对一致性
- 4.10 光谱图评估重复性
- 4.11 由MAQC(生物芯片质量控制)得到的归一化前AffymetriX数据的盒式图
- 4.12 来自MAQC研究得到的AffymetriX芯片数据的SPM(谱图)
- 4.13 存在批次效应的差异表达基因的强度图
- 5.1 信息丰富的和不提供信息的探针集的探针比较
- 5.2 基因过滤对p值分布的影响
- 5.3 不同过滤技术排除基因的百分比
- 5.4 两种过滤技术的差异
- 5.5 基因过滤技术的分布差别
- 5.6 在聚类中的欧几里得(Euclidean)和皮尔森(Pearson)距离
- 5.7 基于欧几里得和皮尔森距离的ALL数据的分级聚类
- 5.8 分级聚类运算的示意图



## &lt;&lt;基于Affymetrix芯片的基因 &gt;

- 5.9 k均值运算的示意图
  - 5.10 ALL数据的主要成分分析
  - 5.11 ALL数据的谱图
  - 5.12 t-检验的可变性
  - 5.13 t-检验
  - 5.14 不良的t-检验:变异对显著性的影响
  - 5.15  $\lambda=0.75$ 的SAM图
  - 5.16 t分布
  - 5.17 使用大样本资料比较两种差异表达检验的方法(30 vs.30)
  - 5.18 使用小样本资料比较两种差异表达检验的方法(3 vs.3)
  - 5.19 各种交互效应的假设方案
  - 5.20 用GLUCO数据中具有不同表达方式的四个基因解释交互效应
  - 5.21 多种检验校正方法及其如何处理假阳性和假阴性
  - 5.22 ALL数据组中调整过和未调整过的p值
  - 5.23 高维性和过度拟合在分离中的关联
  - 5.24 过度拟合的问题
  - 5.25 嵌套循环交叉验证
  - 5.26 利用PAM基因组合秩次升高
  - 5.27 利用LASSO基因组合秩次升高
  - 5.28 交叉验证中的位点排列
  - 5.29 进行分类的最佳基因数量
  - 5.30 惩罚回归:惩罚的系数关联
  - 5.31 神经网络方案
  - 5.32 支持向量机模型的二维可视框图
  - 5.33 使用MLP包含高秩基因组的GO通路
  - 5.34 利用GSA含有高秩基因组的GO通路
  - 5.35 BioCarta通路
  - 5.36 识别差异表达的染色体区域
  - 6.1 热图
  - 6.2 强度图
  - 6.3 基因列表图
  - 6.4 Venn(维恩)图
  - 6.5 火山图
  - 6.6 MA图
  - 6.7 平滑散点图
  - 6.8 柱状图
  - 6.9 数据组HD的盒图
  - 6.10 小提琴图
  - 6.11 密度图
  - 6.12 系统树图
  - 6.13 重要基因组的GO通路
  - 7.1 药物开发中的基因表达谱
  - 7.2 Fos的剂量反应特征
  - 9.1 二代测序排序可能出现的错误
- 表格目录
- 1.1 双通道ANOVA设计
  - 2.1 AffymetriX探针集的类型和名称

<<基于Affymetrix芯片的基因 >

- 2.2 已经不再使用的AffymetriX探针集和名称
- 2.3 原始AffymetriX探针集的注释级别
- 2.4 产生客户可选择的CDF的规则
- 2.5 基于Ensembl Gene数据库的HG U133 plus 2.0探针的使用
- 3.1 不同样本的RNA产率
- 4.1 背景微小差异的影响
- 5.1 修正p值的计算
- 5.2 分类和假设检验
- 5.3 采用LASSO和PAM选择的重要基因
- 5.4 惩罚回归:基因选择
- 5.5 采用MLP选择的重要基因
- 5.6 采用GSA选择的前5个上调基因组和前5个下调基因组
- BioBox目录
  - 1.1 基因表达芯片
  - 1.2 分子生物学的中心法则
  - 1.3 siRNA
  - 1.4 表型
  - 2.1 剪接变异
  - 2.2 基因
  - 3.1 Northern杂交
  - 3.2 转录因子
  - 3.3 血液
  - 3.4 细胞培养
  - 3.5 X染色体失活:Xist
  - 3.6 凝胶电泳
  - 3.7 生物分析仪进行RNA分析
  - 3.8 RTqPCR(荧光定量PCR)
  - 5.1 管家基因
  - 7.1 生物标志物
  - 7.2 EC50,ED50,IC50,LC50和LD50
  - 7.3 生物标志物和临床意义
  - 7.4 基因表达的意义
- 9.1 表观遗传学的实例:DNA甲基化
- StatsBox目录
  - 1.1 关联的两种解释
  - 3.1 能力
  - 4.1 准度和精度
  - 4.2 贝叶斯统计
  - 4.3 可重复性
  - 4.4 关联假设
  - 5.1 参数,变量,统计
  - 5.2 完全拟合
  - 5.3 有监督和无监督的研究
  - 5.4 重取样技术
  - 5.5 神经网络
  - 5.6 多变量投影方法的步骤
  - 5.7 确定差异表达的步骤

<<基于Affymetrix芯片的基因 >

- 5.8 比值的对数=对数差异
- 5.9 零假设和p值
- 5.10 变异,标准偏差和标准误差
- 5.11 经验贝叶斯方法
- 5.12 显著性水平和能力
- 5.13 参数和非参数检验比较
- 5.14 Explanatory和响应变异
- 5.15 通用线性模型
- 5.16 测量规模
- 5.17 交互反应
- 5.18 规则化或惩罚
- 5.19 敏感性和特异性
- 5.20 多重检验校正步骤
- 5.21 信息并不是越多越好
- 5.22 核心技术
- 5.23 刀切法和自助法

## 章节摘录

Chapter 1 Biological question All experimental work starts in principle with a question. This also applies to the field of molecular biology. A molecular scientist is using a certain technique to answer a specific question such as, “ Does the cell produce more of a given protein when treated in a certain way? ” Questions in molecular biology are indeed regularly focused on specific proteins or genes, often because the applied technique cannot measure more. Gene expression studies that make use of microarrays also start with a biological question. The largest difference to many other molecular biology approaches is, however, the type of question that is being asked. Scientists will typically not run arrays to find out whether the expression of a specific messenger RNA is altered in a certain condition. More often they will focus their question on the treatment or the condition of interest. Centering the question on a biological phenomenon or a treatment has the advantage of allowing the researcher to discover hitherto unknown alterations. On the other hand, it poses the problem that one needs to define when an “ interesting ” alteration occurs.

1.1 Why gene expression? 1.1.1 Biotechnological advancements Research evolves and advances not only through the compilation of knowledge but also through the development of new technologies. Traditionally, researchers were able to measure only a relatively small number of genes at a time. The emergence of microarrays (see BioBox 1.1) now allows scientists to analyze the expression of many genes in a single experiment quickly and efficiently. 1.1.2 Biological relevance Living organisms contain information on how to develop its form and structure and how to build the tools that are responsible for all biological processes that need to be carried out by the organism. This information ? the genetic .....

Gene expression microarrays. In microarrays, thousand to million of probes are fixed to a surface, being either glass or silicon chip. The latter explains why microarrays are also often referred to as chips. The target of the probes, the mRNA samples, are labelled with fluorescent dyes and are hybridized to their matching probes. The hybridization intensity, which estimates the relative amount of the target transcripts, can afterwards be measured by the amount of fluorescent emission on their respective spots. There are various microarray platforms differing in array fabrication, the nature and length of the probes, the number of fluorescent dyes that are being used, etc.

BioBox 1.1: Gene expression microarrays content ? is encoded in information units referred to as genes. The whole set of genes of an organism is referred to as its genome. The vast majority of genomes are encoded in the sequence of chemical building blocks made from deoxyribonucleic acid (DNA) and a smaller number of genomes are composed of ribonucleic acid (RNA), e.g., for certain types of viruses. The genetic information is encoded in a specific sequence made from four different nucleotide bases: adenine, cytosine, guanine and thymine. A slightly different composition of building blocks is present in mRNA where the base thymine is replaced by uracil. Genetic information encoding the building plan for proteins is transferred from DNA to mRNA to proteins. The gene sequence can range in length typically between hundreds and thousands of nucleotides up to even millions of bases. The number of genes that contain protein-coding information is expected to be between 25,000 to 30,000 when looking at the human genome. A protein is made by constructing a string of protein building blocks (amino acids). The order of the amino acids in a protein matches the sequence of the nucleotides in the gene. In other words, messenger RNA interconnects DNA and protein, and also has some important practical advantages compared to both DNA and proteins (see BioBox 1.2). Increasing our knowledge about the dynamics of the genome as manifested in the alterations in gene expression of a cell upon treatment, disease, development or other external stimuli, should enable us to transform this knowledge into better tools for the diagnosis and treatment of diseases. DNA is made of two strands forming together a chemical structure that is called “ double helix. ” The two strands are connected with one another via pairs of bases that form hydrogen bonds between both strands. Such pairing of so-called “ complementary ” bases occurs only between certain pairs. ....

Central dogma of molecular biology. The dogma of molecular biology explains how the information to build proteins is transferred in living organisms. The general flow of biological information (green arrows) has three major components: (1) DNA to DNA (replication) occurs in the cell nucleus (drawn in yellow) prior to cell division, (2) DNA to mRNA (transcription) takes place whenever the cell (drawn in light red) needs to make a protein (drawn as a chain of red dots), and (3) mRNA

## &lt;&lt;基于Affymetrix芯片的基因 &gt;

toproteins(translation)istheactualproteinsynthesisstepintheribosomes(drawinggreen).Besidesthese general transferst hatoccur normally in most cells, there are also some special information transferst that are known to occur in some viruses or in a laboratory experimental setting. BioBox 1.2: Central dogma of molecular biology

.....Hydrogen bonds can be formed between cytosine and guanine or between adenine and thymine. The pairing of the two strands occurs in a process called “ hybridization. ” Compared to DNA, mRNA is more dynamic and less redundant. The information that is encoded in the DNA is made available for processing in a step called “ gene expression ” or “ transcription. ” Gene expression is a highly complex and tightly regulated process by which a working copy of the original sequence information is made. This allows a cell to respond dynamically both to environmental stimuli and to its own changing needs, while DNA is relatively invariable. Furthermore, as mRNA constitutes only the expressed part of the DNA, it focuses more directly on processes underlying biological activity. This filtering is convenient as the functionality of most DNA sequences is irrelevant for the study at hand. Compared to proteins, mRNA is much more measurable. Proteins are 3D conglomerates of multiple molecules and cannot benefit from the hybridising nature of the base pairs in the 2D, single molecule, structure of mRNA and DNA. Furthermore, proteins are very unstable due to denaturation, and cannot be preserved even with very laborious methods for sample extraction and storage. When using microarrays to study alterations in gene expression, people normally will only want to study the types of RNA that code for proteins? the messenger RNA (mRNA). It is however important to keep in mind that RNA not only contains mRNA? a copy of a section of the genomic DNA carrying the information of how to build proteins. Besides the code for the synthesis of ribosomal RNA, there are other non-coding genes that, e.g., contain information for the synthesis of RNA molecules. These RNAs have different functions that range from enzymatic activities to regulating transcription of mRNAs and translation of mRNA sequences to proteins. The numbers of these functional RNAs that are encoded in the genome are not known. Initial studies looking at the overall transcriptional activity along the DNA are predicting that the number will most likely be larger than the number of protein-coding genes. People used to say that a large portion of the genomic information encoded in the DNA are useless ( “ junk DNA ” ). Over the last years scientific evidence has accumulated that a large proportion of the genome is being transcribed into RNAs of which a small portion constitutes messenger RNAs. All these other non-coding RNAs are divided into two main groups depending on their size. While short RNAs are defined to have sizes below 200 bases, the long RNAs are thought to be mere precursors for the generation of small RNAs, of which the function is currently still unknown? in contrast to the known small RNAs such as microRNAs or siRNAs[6] (see BioBox 1.3 for an overview of different types of RNA). Microarrays are also being made to study differences in abundance of these kinds of RNA.

.....RNA. In contrast to mRNA (messenger RNA) which contains the information of how to assemble a protein, there are also different types of non-coding RNA (sometimes abbreviated as ncRNA) a. Here are the types that are most relevant in the context of this book: miRNA in length, which regulate gene expression. long ncRNA (long non-coding RNA) are long RNA molecules that perform regulatory roles. An example is XIST, which can also be used for data quality control to identify the gender of a subject (see BioBox 3.5). rRNA (ribosomal RNA) are long RNA molecules that make up the central component of the ribosome. They are responsible for decoding mRNA into amino acids and are used for RNA quality control purposes (see Section 3.1.2.8). siRNA (small interfering RNA) are small double-stranded RNA molecules of about 20-25 nucleotides in length and play a variety of roles in biology. The most commonly known function is a process called RNA interference (RNAi). In this process siRNAs interfere with the expression of a specific gene, leading to down regulation of the synthesis of new protein encoded by that gene. tRNA (transfer RNA) are small single-stranded RNA molecules of about 74-95 nucleotides in length, which transfer a single amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis. Each type of tRNA molecule can be attached to only one type of amino acid. a Non-coding RNA refer to RNA molecules that are transcribed from DNA but not translated into protein. b Ribosomes can be seen as the protein manufacturing machinery of all living cells. c There are, however, also processes known as small RNA-induced gene activation whereby double-stranded RNA target gene promoters to induce transcriptional activation of associated genes. BioBox 1.3: siRNA

.....In this book we will focus on studying mRNA. However, most likely many remarks given

on the experimental design and the data analysis will apply to the study of small RNA as well. 1.2 Research question

The key to optimal data analysis lies in a clear formulation of the research question. Being aware of having to define what one considers to be a “ relevant ” finding in the data analysis step will help in asking the right question and in designing the experiment properly so that the question can really be answered. A well-thought-out and focused research question leads directly into hypotheses, which are both testable and measurable by proposed experiments. Furthermore, a well-formulated hypothesis helps to choose the most appropriate test statistic out of the plethora of available statistical procedures and helps to set up the design of the study in a carefully considered manner. To formulate the right question, one needs to disentangle the research topic into testable hypotheses and to put it in a wider framework to reflect on potentially confounding factors. Some of the most commonly used study designs in microarray research will be introduced here by means of real-life examples. For each type of study, research questions are formulated and example datasets described. These datasets will be used throughout the book to illustrate some technical and statistical issues.

### 1.2.1 Correlational vs. experimental research

Microarray research can either be correlational or experimental. In correlational research, scientists generally do not apply a treatment or stimulus to provoke an effect on, e.g., gene expression (influence variables), but measure them and look for correlations with mRNA (see StatsBox 1.1). A typical example are cohort studies, where individuals of populations with specific characteristics (like diseased patients and healthy controls) are sampled and analysed. In experimental research, scientists manipulate certain variables (e.g., apply a compound to a cell line) and then measure the effects of this manipulation on mRNA. Experiments are designed studies where individuals are assigned to specifically chosen conditions, and mRNA is afterwards collected and compared. It is important to comprehend that only experimental data can conclusively demonstrate causal relations between variables. For example, if we found that a certain treatment A affects the expression levels of gene X, then we can conclude that treatment A influences the expression of gene X. Data from

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:<http://www.tushu007.com>