**生物芯片重点**

**第一章**

1、简述生物芯片产生的历史背景（人类基因组计划）？

答：**人类基因组计划的提出，以及组学的兴起**。

**HGP计划：**由美国科学家于1985年率先提出、于1990年正式启动的、旨在拟在15年内至少投入30亿美元，进行对人类全基因组的分析。发现所有人类基因并搞清其在染色体上的位置，破译人类全部遗传信息的科学计划。美国、英国、法国、德国、日本和我国科学家共同参与了这一计划。

**HGP预期目标：**鉴定约 80,000 到140,000 个人类基因；测定人类基因组30亿个碱基的序列；贮存所有的人类基因组序列信息；开发一系列信息学分析工具；解决一系列伦理 、道德和法律问题；研发相应的组学研究技术平台

**HGP完成目标：**人类基因组由30亿个碱基组成；人类基因的平均大小为3000个碱基,但基因间差异巨大（最大的基因为240万个碱基组成的dystrophin基因）；人类基因数为250,00-30,000个——远低于预期的估计；人类个体之间约99.9%的碱基序列组成相同；超过50%基因的功能未知

2、什么叫生物芯片？包括哪些具体类型？有哪些优缺点？

答：

**生物芯片：**指能储存大量生物信息或快速并行处理多个生物样品的微器件，它的加工运用了微电子工业中十分成熟的光学光刻技术和微机电系统加工中所采用的各种方法，所处理的对象是生物样品，故名。

**类型：**

**1)基因芯片：**又称DNA芯片(DNAchip)、DNA微阵列(DNAmicroarray)是将大量DNA片段通过一定方式固定于某种固相载体表面，形成致密、有序的DNA分子点阵。

**2)蛋白芯片：**是指固定于支持介质上的蛋白质（抗原、抗体、小肽、受体和配体、蛋白质-DNA和蛋白质-RNA复合物等）构成的微阵列。

**3)细胞芯片：**将不同DNA探针（能转染的脂质体）点样在玻璃片上，做成DNA微阵列芯片，然后在DNA微阵列芯片上培养哺乳动物细胞，在芯片上的DNA在转染试剂的作用下原位感染细胞，获得带有不同遗传性状表型细胞。

**4)组织芯片：**也称为组织微阵列(tissuemicroarray，TMA)，是指在基质（玻片等）表面固定大量的、可寻址的微小组织样本，用于高通量地检测不同组织中DNA、RNA和蛋白质等分子的变化情况。

**5)芯片实验室（Lab-on-chip)：**是生物芯片技术发展的最高阶段。它是一种高度集成化的芯片，是集样品制备、基因扩增、核酸标记及检测为一体的便携式生物分析系统。

**优缺点：**

**优点：**可以同时进行多种细胞的检测，同时也可以获取待测样品的信息。

**缺点：**不能直接分析细胞内的表达产物，无法反映机体内情况

3、什么叫DNA microarray？它的基本原理与特点是什么？有哪些基本类型？依据探针，载体，用途分为哪些基本类型？

答：

**DNA微阵列(DNA microarray)：**即基因芯片。是将大量DNA片段通过一定方式固定于某种固相载体表面，形成致密、有序的DNA分子点阵。

**基本原理：**待测样品核酸分子经过标记，与固定在载体上的 DNA 阵列上的点按碱基配对原理同时进行杂交。通过激光荧光检测系统等对芯片进行扫描，检测杂交信号而获取样品分子的数量和序列信息，用计算机软件进行数据的比较和分析，从而对基因序列及功能进行大规模、高通量的研究。

**特点**：

高通量、平行化、微量化、自动化、低成本、网络信息化

**类型：**

制备方式: 原位合成芯片、DNA微集阵列

探针类型: cDNA芯片、寡核苷酸芯片

杂交过程：双色杂交芯片，单色杂交芯片

用途：表达谱芯片、诊断芯片、指纹图谱芯片、测序芯片等

4、简述DNA microarray的应用领域包括哪些方面（重点关注医学和生物领域）？

答：

生命科学研究领域：生物进化和分类，生物发育、衰老、分化，信号转导、基因表达调控，生物体对外界刺激的反应，生理生化研究，寻找新基因或基因、新功能新途径

疾病诊断：遗传性疾病诊断、疾病分型诊断、传染性病原体诊断、遗传及重大疾病的早期诊断

医学研究、药物研发及临床应用：基因表达谱差异做发病机制研究、疾病药物作用机制的研究、寻找药物靶位点、筛选先导药物、个体化医疗、中药现代化。

5、生物芯片技术产业的前景及其思考（了解这门技术的现在的发展瓶颈和不足）？

答：

**现存的问题及挑战：**

1)关键技术瓶颈

①提高生物芯片的稳定性；

②样品制备和标记操作简化；

③增加信号检测的灵敏度；

④高度集成化样品制备、基因扩增、核酸标记及检测仪器的研制和开发等。

2)生物信息学挖掘（技术多成熟，大量数据如何进一步发掘）

3)生物芯片标准化

Ps:2002年，第一个生物芯片国际专利。

6、R语言的基本操作，如何安装？安装的方法有哪些？调取数据的方法有哪些？bioconductor安装程序包的方法？

答：

**R开源的数据分析软件：**是用于统计分析、绘图的语言和操作环境。R是一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具。

R语言：开源，有丰富的算法包和强大的画图能力，我们可以通过各种各样的

程序包或者修改源代码来满足我们自己对数据分析和挖掘的需求。

**Rstudio**：开发者基于R开发了相应的一个能够实现菜单化操作的Rstudio。

R的程序包：相当于R的一个插件，可以实现多种不同的功能。

**R包的安装**：install.packages()命令。R包安装的位置：一定要安装到R文件夹下的library文件夹中！

**R包的调取：**library（）

**Bioconductor**：是建立在R语言基础上的，用于生物信息数据的注释、处理分析及可视化工具包的总集，由一系列R程序包组成，也是开源免费的。

**Bioconductor包的安装**：

source(“http://bioconductor.org/biocLite.R”)

biocLite(“package\_name”)

**第二章**

1、简述DNA芯片制备的基本过程？了解杂交点的含义（阴性对照，阳性对照，外源基因…….）及作用？数据表的行列代表什么含义？

答：

DNA芯片制备的基本过程：①探针的设计与制备②芯片设计③基片的制备④探针点样或合成⑤质量控制  
各杂交点的含义和作用：

空白对照：抵消芯片本身的非特异性杂交的干扰信号

阴性对照（点样液）：检测点样液是否有污染，抵消空白样液的干扰信号

阳性对照：检测探针是否可用

外标（属于阳性对照）：衡量操作者的操作水准。将与本次实验无关的外源基因的探针也标记上，和本次实验目的探针混合一起杂交，如果外标点信号较好则表明此次操作合格。

2、DNA芯片设计的基本原则？

答：

**DNA芯片设计包括：**探针本⾝的设计和探针在芯片上的布局两方面。

首先要确定芯片所要检测的目标对象，获得相应的DNA序列数据，找到序列特性作为设计的参考依据，还要获得关于序列的突变信息及其他信息。

其次在设计和布局时需要考虑如下几个方面：  
互补性、敏感性、特异性、可控性、可读性

3、探针有哪些类型（cDNA，寡核苷酸，原位表面化学合成）？定义是什么？设计合成的过程？优缺点？主流芯片厂家采用的是什么探针类型？

**基因芯片探针类型：**

按来源分：  
**①PCR产物探针：**直接利用基因组进行PCR扩增生成的DNA片段作为探针制成的微阵列。适用于未知原核生物及CGH芯片

制备方法：通用引物PCR扩增、基因特异性引物PCR扩增

优势：

（1）可应用于未知基因组序列信息芯片的制备  
（2）PCR产物比较长，产生强烈的荧光信号  
（3）简单易行，成本低  
缺点：

（1）PCR产物为双链核苷酸，使用前需要进行变性处理  
（2）探针序列过长，容易发生非特异性杂交

**②cDNA探针：**采用mRNA反转录生成的cDNA作为探针制成的微阵列被称为cDNA微阵列。主要用于未知序列的真核生物。

制备方法：RT-PCR

优点：适合于某些科学研究，杂交信号好  
缺点：费时费力，成本高，特异性不强

**③寡核苷酸探针**

按生产方式分：  
**①原位合成探针（**原位合成寡核苷酸探针**）：**

(1)Affymetrix公司利用原位光导合成法合成长度一般为25 nt的探针

制备方法：光导原位合成技术：又称为光导化学合成，是Affymetrix公司将固相DNA合成和光刻技术有机地结合在一起，用于在石英片表面合成探针的技术。

优势：  
特异性好： 25mer特异性强  
灵敏度高： PM-MM探针优势  
重复性好：检测值可靠

(2)Agilent公司等通过半导体光刻法或分子印章法合成60 nt 左右的长链寡核苷酸探针

制备方法：目前有分子印章法、原位喷印合成法和半导体光刻法等方法。

**②直接点样的探针**

**③设计合成后再点样探针（**合成后点样寡核苷酸探针**）：**通过特定的生物信息学软件设计针对每个基因的特异性寡核苷酸探针，长度一般为40－100 nt，用DNA合成仪合成后用点样仪点样于处理后的玻片表面。主要适用于序列已知的真核生物基因芯片。主要用于真核生物芯片。

主要使用公司：Clontech公司的Atlas系列、Operon公司的Array-Ready Oligo Sets

设计基本原则：  
A、探针设计在60－70个碱基长度能保持杂交特异性和杂交信号强度的平衡；  
B、尽可能在靠近3′端1.5kb内设计探针（原因和mRNA标记方法有关）；  
C、不同探针间的Tm值相差不超过±5℃；  
D、探针序列和其他基因同源性小于70％；  
E、探针内部本身不含有7个相同的碱基。

|  |  |  |  |
| --- | --- | --- | --- |
| 类型 | 来源 | 优点 | 缺点 |
| cDNA探针 | mRNA逆转录 | 经济 | 操作繁琐、特异性不强 |
| PCR扩增探针 | 基因组PCR扩增 | 经济 | 操作繁琐、限于原核生物 |
| 长链Oligos探针 | DNA合成 | 方便、容易操作 | 费用高 |
| AffyPM-MM探针 | DNA合成 | 数据质量高 | 费用高、专利限制 |

4、探针和基片如何连接？要注意什么？

答：

**接触式点样法和非接触式点样法**

接触式点样法：点样工作头或点样工作头中的液体和基片间有直接接触的各种方法，是目前微阵列生产中最常用的方法。

非接触式点样:不需要点样工作头和微阵列表面直接接触的技术，一般通过微量喷墨印刷来实现。

**点样后处理**

①紫外交联增强点样区探针与活性基团的连接

②含SDS 的溶液洗去未结合的探针

③封闭剂封闭未与探针结合的游离活性基团

5、简述玻片的表面化学处理方法及相应的探针固定方法？玻片的制备质量标准有哪些？

答：

**表面化学处理：**片基的表面化学修饰过程，即采用不同活化试剂通过化学反应在基片表面衍生出氨基、醛基、异硫氰酸基及环氧基团等活性基团，通过与DNA分子中的磷酸基、氨基、羟基等基团形成离子键或共价键。

作用：  
• 固定探针，防止探针在杂交时被洗脱  
• 提高探针结合能力，增强杂交信号  
• 减少探针扩散，提高探针密度  
**探针固定方法：**

①氨基修饰：使玻片的表面覆盖一层有活性的氨基基团(一NH2)。

过程：第一步：化学清洁剂处理；第二步氨基硅烷化试剂对玻璃表面进行修饰，对非氨基基团进行封闭；第三步离子键固定作用

优点：表现出较低的自发荧光（杂交前）和低的荧光背景（杂交后）；适用于PCR产物、cDNA和蛋白质等大分子为探针芯片类型。  
缺点：非特异性偶联；DNA分子还必须在杂交前失活

改进：多元胺的网格化玻片

②醛基化修饰

③异硫氰基修饰

④二硫键修饰

 玻片的制备质量标准：

可结合性：与探针是否高效结合，玻片化学表面结合探针分子的能力和效率。

耐受性：表面涂层是否稳定，基片化学表面微阵列分析过程中的稳定性。常用标准是在整个芯片的检测实验过程中，整个探针分子的损失不超过10％。

均一性：基团分布是否均匀，芯片基片表面化学活性基团的密度。均一表面是芯片表面具有相同数目的反应活性基团，即相同的活性基团密度，一个通用的标准是整个芯片基片表面的偏差不超过25％。

荧光惰性：背景噪音是否很低，表面能牢固结合样品中的靶标分子，但不会引起信号的系统性偏差，称为具有荧光惰性。

平整度：基片表面是否平整，芯片局部表面的二维尺度的平整性。

尺寸规格：尺寸大小是否合适，基片材料物理尺寸的评价标准，两种规格：25mm×76mm×0.94mm12.5mm×12.5mm×0.94mm长宽允许的误差范围是300微米和厚度的误差范围是10微米

经济性：处理成本是否低廉

6、基因芯片的质量评价标准包含哪些方面？

费用（**affordability**）：芯片制备的成本

容量（**content**）：芯片表面所包含的基因等生物材料的总量。

密度（**density**）：基片单位面积内的探针数，芯片密度（点数/cm2）可以通过测量两相邻点的中心间距（centertocenterspacing,CTC）来计算密度。密度的大小决定了在单位面积上能获得的数据量的多少。一般的接触式和非接触式点样的密度为1000－10000点数/cm2，而原位合成可以实现250000点数/cm2。

样品点尺度（**featuresize**）：微阵列中样品点（**feature**）的大小，接触式和非接触式点样的样品点尺度为75－300微米，而原位合成技术样品点尺度为10－40微米。样品点越小，最大密度越大，最大密度越大，单位面积上获得的信息越多，所以制备的样品点尺度越小，微阵列制备仪器的价值越高。

样品点纯度（**purity**）：芯片中各个样品点中所包含的分子的纯度，样品点分子纯度的指标一般为**99**％。纯度决定反应的特异性，杂质造成的非特异荧光会使得后续信号及数据分析变得很复杂。

样品点活性（**featurereactivity**）：探针点分子的活性或者有效性，热、酶或者化学破坏引起。大多数的阵列探针活性可以达到**50**％以上。

规整度（**regularity**）：微阵列上各点所在行列的对齐程度，规整度通常用样品点间中心距离的差异来衡量，一般用点间距偏差10％。规整度的偏差有多个来源，包括基片表面、制备仪器和制备环境等。规整度是一个重要指标，因为它决定从图像中提取数据的难易程度。

可操作性（**easeofimplementation**）：芯片制备技术在实验室被掌握和使用的难易程度

通量（**throughout**）：单位时间加工微阵列的数目，一般以每天能加工制备的芯片数目表示。

**第三章**

1、表达谱基因芯片实验的基本过程？如何针对性地设计实验（差异基因的筛选，信号通路的筛选，针对时间序列的筛选，肿瘤的分型诊断）（综合题）（单因子，双因子，时间序列，多因子实验设计）？

答：  
表达谱基因芯片实验的基本过程：芯片制备、样品制备、杂交反应和信号检测和结果分析。

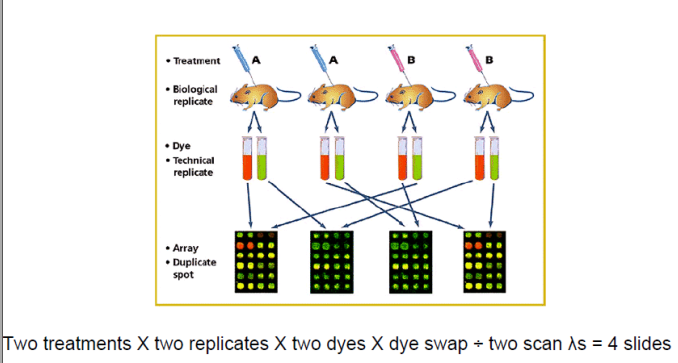
**1.芯片制备**  
目前制备芯片主要以玻璃片或硅片为载体，采用原位合成和微矩阵的方法将cDNA作为探针按顺序排列在载体上。芯片的制备除了用到微加工工艺外，还需要使用机器人技术。以便能快速、准确地将探针放置到芯片上的指定位置。

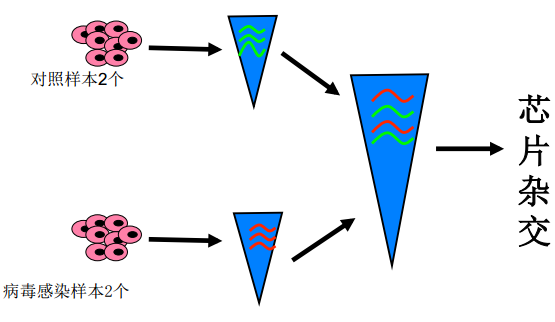
**2.样品制备**  
生物样品往往是复杂的生物分子混合体，除少数特殊样品外，一般不能直接与芯片反应，有时样品的量很小。所以，必须将样品进行提取、扩增，获取其中的mRNA反转录得到的cDNA，然后用荧光标记，以提高检测的灵敏度和使用者的安全性。

**3.杂交反应**  
杂交反应是荧光标记的样品与芯片上的探针进行的反应产生一系列信息的过程。选择合适的反应条件能使生物分子间反应处于最佳状况中，减少生物分子之间的错配率。

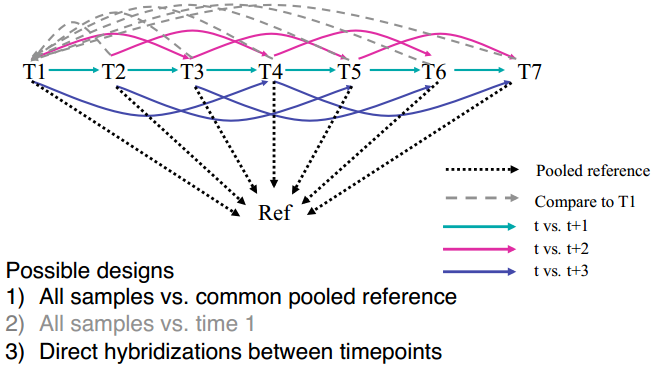
**4.信号检测和结果分析**  
杂交反应后的芯片上各个反应点的荧光位置、荧光强弱经过芯片扫描仪和相关软件可以分析图像，将荧光转换成数据，即可以获得有关生物信息。

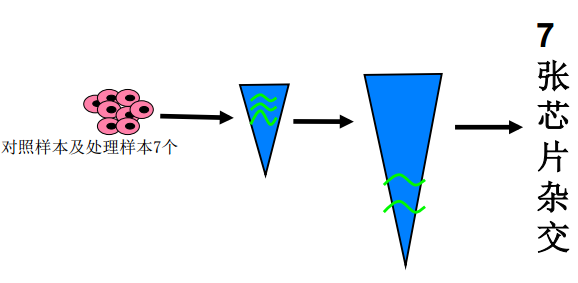
**①单因子两组间差异比较实验设计（差异基因的筛选，肿瘤分型诊断等）：**

****

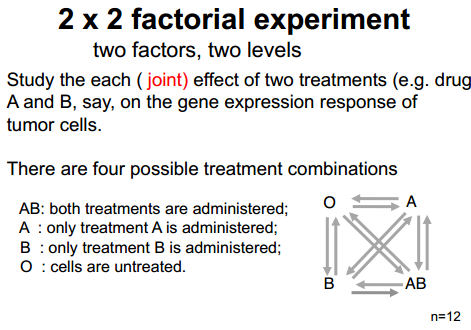
****

**②单因子连续型差异比较实验设计（时间序列等）：方法： 聚类分析**

****

****

**③两个因子间差异比较的实验设计：**



2、如何获得mRNA的样本？如何判断样品是否合格？样本标记的基本方法有哪些？原理？优缺点？

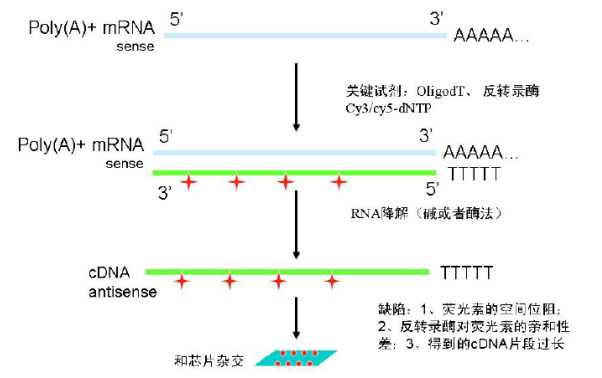
答：

获取：Trizol法总RNA提取： 去rna酶，保持低温

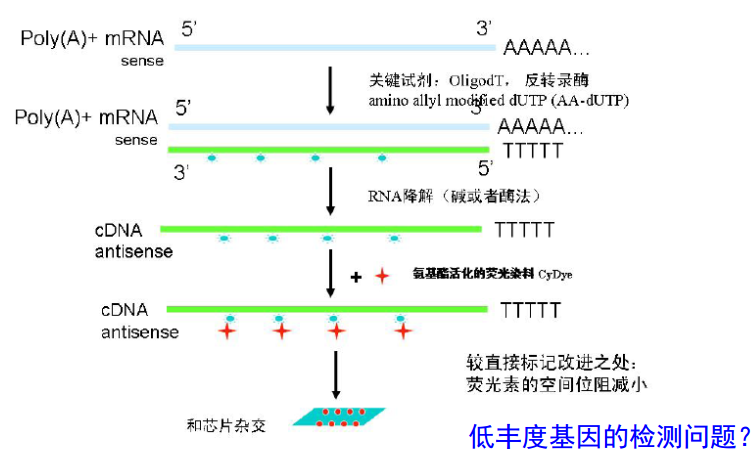
判断是否合格：测定样品浓度，检测片段大小，RIN和28S:18S的方法

样本标记的基本方法：

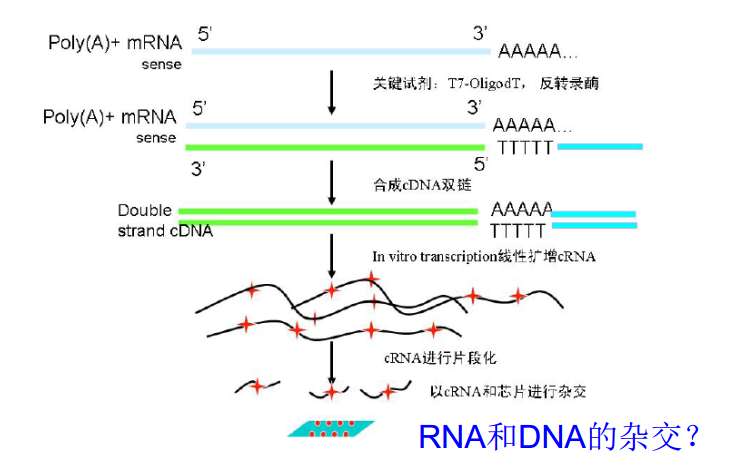
1、直接逆转录标记



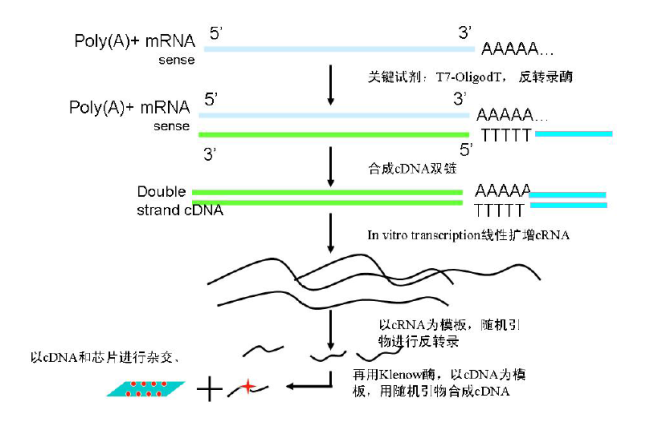
2、间接标记法



3、RNA聚合酶体外转录线性扩增法



4、改良的RNA扩增标记方法

****

3、杂交的过程？要注意什么影响因素？如何判断是否成功？

答：

•预杂交：封闭和失活氨基或醛基；洗掉未结合的探针

•杂交：一般把标记好的总RNA样本溶解在12μl-80μl杂交液中，然后盖上盖玻片在水浴中恒温静止或者动态杂交过夜。

探针序列；

杂交温度（选择适当的杂交温度是基因芯片核酸分子杂交成败的关键因素之一, Tm值是关键。杂交温度 在低于Tm值20～25℃温度下进行）；

探针浓度；靶基因浓度；

杂交液组分（盐：高离子强度溶液中，其正离子可中和DNA链磷酸基团的负电荷，消除其间的静电斥力，有利于杂交分子的形成。一般杂交体系的离子强度为5×SSC或6×SSC。甲酰胺（DMSO））

•洗脱干燥

4、芯片扫描的基本原理与方法有哪些？什么时候用？过程？优缺点？

|  |  |  |
| --- | --- | --- |
|  | 非共聚焦扫描 | 共聚焦扫描 |
| 原理 | CCD(电荷偶合装置) | PMT(光学倍增管) |
| 光源 | 汞灯 | 激光 |
| 检测方式 | 大面积扫描 | 单点检测 |
| 扫描速度 | 快， 0.5-2分钟 | 慢， 10分钟以上 |
| 检测质量 | 误差大， 灵敏度低 | 精度高 |
| 用途 | 教学等低密度芯片 | 高密度检测 |

5、图像分析的基本过程是什么？如何判断一副图像是否合格？

答：

**基本过程：**

图像质量评估

图像数据提取与计算：

划格（**gridding**）和定位（**addressing**）是将事先根据芯片型号定义好行列数的格子覆盖到芯片上，确定样点的位置。

分割（**Segmentationorspotpicking**）通过一系列背景计算方法，将杂交的荧光信号像素与背景像素分开

信息提取（**intensityextraction**）包括计算荧光信号强度和背景强度，还要将背景扣除。

数据质量评估（**qualitymeasurement**）根据统计数据来评估芯片或者样点质量。

**判断图像合格的依据：**信号点大小形状是否规则、信号点位置是否正确、信号点密度是否足够、信号点的变异程度、背景变异程度、信号点的饱和度。

6、杂交和背景信号的提取与分析方法有哪些？

答：

**提取分析方法：**

**①定位方法：**

手动定位、半自动定位、自动定位

**②分割方法：**

a 基于像素空间位置的分割方法、b 基于像素强度的分割方法、c 综合像素强度和空间信息的分割方法

**③前景和背景信号值的计算：**

背景计算方法: 局部法、亚栅格法、对照基因法、空白对照法。

7、如何理解芯片杂交信号与基因表达的关系？

答：某单色信号越强即其所代表的基因表达越强。

8、单色系统和双色系统的共同点和差别？

答：

共同点：均可用于基因表达差异的分析。

不同点：双色系统运用的是cy3/cy5比值可作为同一张芯片上不同基因表达差异的标记；单色系统运用的是生物素标记的不同单张芯片之间比较。

**第四章**

1、为什么要均一化？针对什么样的系统偏差？（染料，针头……）如何均一化？怎样针对不同的系统偏差进行均一化？

答：

**均一化目的：**为了移除基因表达谱数据中存在的系统误差, 包括：样本制备、杂交方法、空间位置、扫描仪器设置、实验操作者等

**均一化方法:**

（1）以所有基因作为均一化参照基因  
（2）看家基因作为均一化参照基因  
（3）对照或者异源基因作为均一化参照基因  
（4）大量的按秩排列的基因作为均一化参照基因  
Within array normalization： Location normalization  
 1. Global normalization  
 2. Intensity dependant methods：linear regression；LOWESS；LOESS  
 3. Print Tip Normalization  
Paired Slide Normalization: dye swap  
Between array normalization： Scale normalization

2、基因芯片数据预处理包括哪些方面？

答：

1）对数转换：生物学上易于理解和解释、满足常用统计方法的前提和使用的方便性  
2）重复值的处理：不一致重复值的去除以及重复值的合并  
3）弱信号的处理：阈值法、信噪比法、分布函数法  
4）缺失值的处理：基因芯片表达谱数据通常会有很多的缺失值，很多统计学方法，如： 聚类分析法和主成分分析，法无法对有缺失值的数据进行分析解决这一问题的两种方案：①除去那些有大量缺失值数据的基因②输入一个缺失值的替代值  
5）基因表达模式过滤  
6）基因表达模式的标准化  
7）非目标基因的去除

3、基因芯片芯片数据系统误差来源有哪些？

答：

1）染料原因：不同荧光素的量子光学差异、不同荧光素的整合效率、不同染料的杂交效率

2）相同样本不同标记方法、不同玻片、不同芯片平台等

3）相同打印仪不同打印组间差别

4）不同杂交体系间结果差异

5）探针的类型、基片的类型

6）批次基片的处理方法、时间、表面特性、空间位置

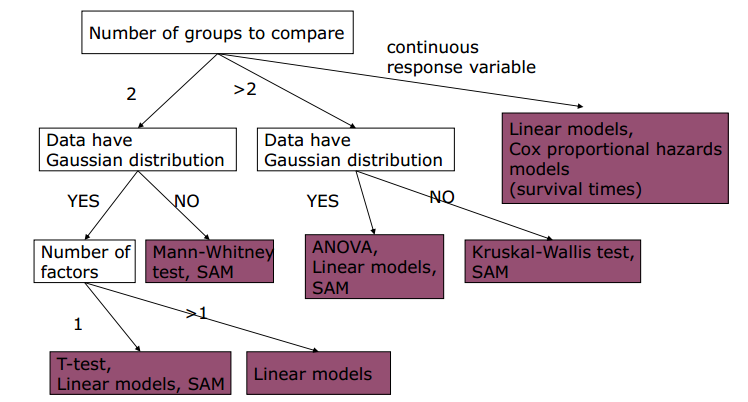
7）样本的提取方法、标记方法

8）不同的分析软件及数据提取方法

9)不同的操作技术员

**第五章**

1、实验设计方案不同，差异基因筛选方法不同（双因子——t检验，多因子——方差分析），如何选择？

****

2、差异基因筛选方法？（均值法，z值法，t检验，anova），t检验和均值法结合使用。T检验和方差分析的结果的含义？P值什么含义？方差f值什么含义？如何检验？

答：

**T 检验结果常与倍数法结果一起使用**：  
• 在 p<0.05显著性水平下, 某个基因表达上调或者表达下调2.0倍（或者1.5倍）；  
• 在 p<0.01或者0.001显著性水平下某个基因表达上调或者表达下调2.0倍（或者1.5倍）；

**F值理解：ANOVA法检验**

F值越大，越说明总的方差波动中，考察的因素方差是主要部分；   
F值越小，说明考察的因素方差与随机因素相比没什么区别。

**p值理解：**某个基因表达上调或者表达下调的可能性小于p

3、验证的基本方法？（rt-pcr,wb……）

答：Northern blotting、半定量RT-PCR、荧光定量RT-PCR、Western blot

**第六章**

1、什么是聚类分析？

答：聚类分析：按照基因表达数据间的相似性将大量的基因芯片表达谱数据聚集成多个小类。 在同一个类内对象之间具有较高的相似度， 不同类之间的对象差别较大。

2、聚类分析法有哪几种常见类型？

答：系统聚类法 （hierarchical method）、分割方法 （partitioning method）、基于密度的方法（density-based method）、基于网格的方法（grid-based method）、基于模型的方法（model-based method）

3、聚类分析的图纵行横行什么意思，颜色代表的什么意思？

答：基因表达谱的目标在于通过基因表达寻找相同模式的基因，这些基因在功能上可能相似或者相关。基因表达谱数据用于聚类分析的结果以一种矩阵的方式表现出来，行代表基因、 列代表样本、每个单元代表基因的绝对或相对量值（LOG值或者比值）

4、表达谱芯片数据的聚类分析有哪些基本应用（信号通路筛选，临床诊断分型，时间序列分析）？

答：差异基因筛选，信号通路筛选，疾病临床诊断分型，时间序列分析等

**第七章**

1、GEO，GO，KEGG等基本数据库的基本用途，在基因注释，功能分析的应用特点？各自的优势？

答：

GO数据库:形成一套描述分子生物功能的词汇, 能够运用于所有的模式生物；便于不同研究机构、组织或研究者交流使用；提高不同数据系统间的兼容性；用于大规模功能基因组学技术来发现功能相似的差异表达基因。

包括：

1）Biological process:Cellular events to which the gene product contributes  
2）Molecular function:Gene product at biochemical level  
3）Cellular component:Location or complex of a gene/protein

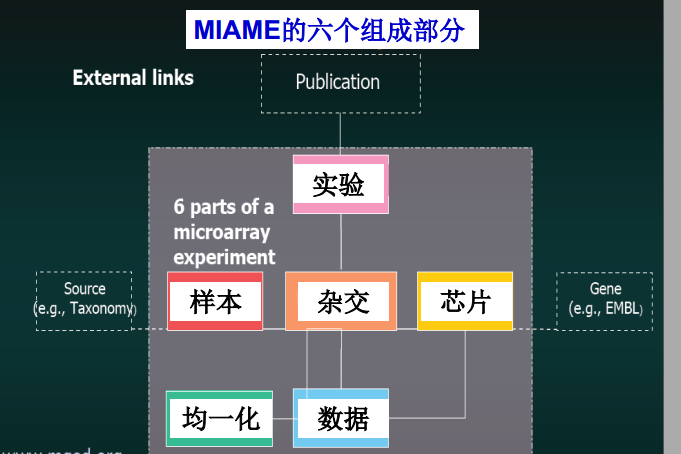
**KEGG数据库：**主要研究代谢通路

**GEO数据库：**基因芯片数据库，含有芯片的平台、样本、系列数据等信息。

2、数据标准化的意义和重要性？怎样进行数据提交？向GEO提交数据的标准？如何提交？提交什么内容？GEO数据库有些什么信息？从GEO下载数据包的基本操作？

答：

标准化的意义：便于同行间的阅读、评议与使用；建立基于**GO**的便于查询的数据库；很多杂志发表的要求。

提交数据的标准：**MIAME：**数据使用者可以不必要知道有关芯片实验数据的详细信息进行分析；芯片实验数据可以用来进行自动与规模化数据分析与挖掘；数据来源于不同的实验室或技术平台。

向GEO提交数据的方法：

第一步 在NCBI上**注册账号**；

第二步 点击**“submit to GEO”**进入提交数据通道；

第三步 **选择芯片的类型**；

第四步 **准备提交数据**：对于AFFY芯片，我们需要准备的是一张**metadata表和processed表（即matrix表），原始数据（cel格式的文件）**；

第五步 **将原始数据和表格一起打包并命名**，以便GEO客服更快的找到您的数据进行审核；

第六步 **回到之前“GEOarchive”界面**，点**击“submit”选择压缩文件，上传类型，以及释放日期**；等待完成即可；

第七步 **给GEO客服写一封信**，让他们知道你上传了数据，并让他们尽快发布。信件内容：说明此次上传的**芯片数据类型，存放的压缩包名字，上传的账户，以及压缩包里包含的数据内容**。

GEO数据库的信息：

①提交者：提交者的联系和登录信息。与许多平台，许多样本和许多系列有关系。

②平台：关于用于以高通量方式检查样本的物理试剂的信息。与一个提交者，许多样本有关。

③样本：关于被检查的mRNA样本，实验条件，和实验产生的基因表达测量数据信息。与一个提交者，一个平台和许多系列有关。

④系列：样本收集，样本是如何相关的，如何排序的，分析是如何进行的，和聚类数据是如何获得的信息。与一个提交者，许多样本有关。