

1、如何理解 G 显带技术在染色体病分析中的金标准地位？基因组学的发展如何推动染色体病的分析和研究？

① 理解 G 显带技术在染色体病分析中的金标准地位：

G 显带技术是一种广泛应用于细胞遗传学研究的染色体显带方法，通过对染色体进行染色、以形成可见的条带模式，从而生成核型图谱，用于分析染色体异常。这一技术利用了染色体中异染色质和常染色质的结构与功能差异，结合染色体的物理化学性质，通过 **Giemsa** 染色产生深浅不一的条带，从而精确识别染色体的特定区域及其异常。

G 显带技术的操作步骤如下：首先，从细胞分裂中期的染色体中提取样本，并利用胰蛋白酶对染色体进行部分消化，以去除染色体表面的蛋白质并暴露出其内部结构。随后，应用 **Giemsa** 染料对染色体进行染色。经过染色处理后，由于异染色质与常染色质会呈现出深浅不同的条带特征：富含腺嘌呤（**A**）和胸腺嘧啶（**T**）的异染色质区域因其高度浓缩的结构和低转录活性，与 **Giemsa** 染料的关键成分具有更强的结合能力，导致异染色质区域的染色较深；相对而言，富含鸟嘌呤（**G**）和胞嘧啶（**C**）的常染色质区域更为开放，染色质结构松散且转录活性较高，DNA 更多地与染色体蛋白（如组蛋白）紧密结合，使得 **Giemsa** 染料较难有效附着，因而染色较浅。完成染色后，通过显微镜可以清晰地观察到染色体上深浅相间的条带。根据染色体的形态特征，我们可以从染色体的着丝粒向端粒端逐步对这些条带进行编号，各条染色体独特的条带模式可为染色体区域的精确定位和描述提供重要依据。

在染色体病的分析中，G 显带技术能通过染色体的条带模式来检测染色体的结构异常和数目变化，为遗传病的诊断提供重要依据。在染色体结构异常方面，G 显带能检测染色体的缺失、重复、倒位、易位等异常。若某条染色体的条带缺失或重复，可能提示与该区域相关的遗传病，如 **22** 号染色体长臂的 **22q11** 缺失综合征；此外，G 显带还能通过对条带模式的分析，检测染色体间的相互易位，如检测慢性粒细胞白血病中典型的费城染色体（**9** 号和 **22** 号染色体间的易位）；而在染色体数目异常的分析中，G 显带能够准确识别染色体的增减，如三体综合征和单体综合征。

通过对染色体数量和条带模式的分析，G 显带技术为染色体病的诊断提供

了直观而可靠的依据。作为一种经典的染色体分析方法，G 显带的优势在于其操作简便、实验设施要求低，同时能够在细胞分裂中期实现全组染色体的可视化分析。相比之下，其他显带技术和染色体分析方法在特定应用场景下虽然具有独特的优势，但在普遍性或操作可行性方面不及 G 显带：Q 显带利用磷酸或柠檬酸缓冲液处理染色体，并结合荧光染料 **quinacrine** 进行染色后，通过荧光显微镜观察荧光条带。由于其实验操作较复杂且需要专业设备，使用范围相对局限；R 显带作为 G 显带的反向显带技术，通过热磷酸缓冲液处理染色体，可突出显示富含 GC 序列的常染色质区域。但与 G 显带相比，其应用也更为特定；而其他显带技术如 C 显带、T 显带和 N 显带则主要针对染色体的特定区域进行分析——C 显带专门显色染色体的结构性异染色质部分，T 显带聚焦染色体端粒区域，而 N 显带用于标记核仁组织区。相比经典的显带技术，

荧光原位杂交（**FISH**）技术具有较高的分辨率，可通过特异性探针检测靶向基因或染色体片段的异常。然而，**FISH** 的检测范围局限于预先设计的目标区域，难以对全基因组进行全面筛查；光谱核型分析（**SKY**）则以染色体的整体标记为基础，精确检测染色体之间的复杂重组和易位，但由于分辨率较低，对单条染色体内部的微小结构变异（如小缺失或小重复）不够敏感；与之形成对比的是，比较基因组杂交（**CGH**）技术可覆盖全基因组水平，检测拷贝数变异（**CNV**）如基因扩增或缺失，但其局限于 DNA 拷贝数的分析，无法提供染色体结构变异的具体信息。

因此，G 显带作为一种能够在普通光学显微镜下直观观察整个染色体结构和条带模式的技术，具有广泛的适用性和较低的实验门槛，且在大多数染色体病的诊断中，G 显带的分辨率已足以满足检测要求。因此，尽管随着技术进步出现了许多高分辨率的分子检测手段，G 显带仍然在染色体病的初步筛查和诊断中占据重要地位，是目前染色体病分析中最具可行性和普遍性的检测方法，是染色体病分析中的金标准。

② 基因组学的发展如何推动染色体病的分析和研究：

基因组学技术的发展经历了从一代测序到三代测序的技术迭代与其他非测序技术的创新，它们为染色体病的研究提供了多层次的工具支持。

(1) 基因组学测序技术发展推动染色体病研究：基因组学测序技术的持续发

展极大地推动了染色体病的研究与诊断。一代测序（**Sanger** 测序）作为基因组学研究的奠基技术，以其高准确性成为早期基因定位和突变鉴定的主要工具，为染色体病的基础研究奠定了重要基础。然而，由于其低通量和高成本，一代测序难以满足全基因组染色体病研究的需求。二代测序（**NGS**）技术的兴起，使得基因组研究进入高通量测序时代。通过快速测序全基因组，二代测序能够精准识别点突变、小片段插入和缺失等基因组变异，这为染色体病中基因突变的全面解析提供了强有力支持。二代测序也可以检测大规模染色体变异（如倒位、重排和拷贝数变异），但这一过程主要依赖算法对测序数据进行推断分析，而非直接观察到变异结构。这种间接性导致其在处理高重复序列或复杂变异时存在一定的局限。三代测序技术以其长读长测序的特点克服了上述不足。单分子实时测序（**SMRT**）和纳米孔测序技术能够直接解析大规模染色体结构变异，特别是在高重复序列和复杂基因组重排的解析上表现出显著优势。这为全面揭示染色体病的复杂变异模式提供了重要信息。此外，传统测序技术通常基于组织层面数据，难以区分不同细胞类型的基因组变异。而单细胞基因组测序技术的出现，则实现了在单细胞水平上解析染色体病的细胞异质性，从而为研究染色体病的动态变化与病理过程提供了更高的分辨率和更精细的分析。基因组测序技术的不断演进显著提高了染色体病研究的深度和广度，为临床诊断和个性化治疗开辟了新的路径。

（2）基因组学非测序技术创新推动染色体病研究：在基因组学测序技术不断推动染色体病研究的同时，非测序分析技术的创新同样为染色体病的研究和诊断提供了重要支撑。这些技术通过解析基因组的结构和功能，从多个维度补充了测序方法的局限性。基因芯片技术通过探针与靶标杂交，可以高效检测特定区域的基因表达、**CNV** 和单核苷酸多态性，为染色体异常的快速筛查提供了高通量的手段；**FISH** 技术利用荧光标记探针靶向特定的基因或染色体区域，能够直观检测染色体的易位、缺失和重复等异常。其高灵敏度和特异性使其在染色体重排、基因定位和癌症基因诊断中发挥了不可替代的作用。基因芯片和 **FISH** 技术均局限于预设目标区域，难以实现全基因组水平的异常检测，而 **CGH** 通过荧光标记样本和对照基因组的杂交，从而在全基因组范围内检测拷贝数变异。这种技术显著提高了染色体扩增和缺失的检测精度，但对于染色体结

构变异（如倒位和易位）的解析能力较弱；针对这一问题，光学图谱技术通过长片段 DNA 分子的荧光标记模式，从而直观检测包括倒位、易位和缺失在内的大规模染色体结构变异。这一技术适用于复杂基因组的解析和研究，是对测序技术的重要补充。这些非测序技术的综合应用，为染色体病研究提供了从单基因水平到全基因组层面的多维度视角，不仅提高了检测的灵敏度和精确度，还推动了对染色体病分子机制的深入理解，为诊断和治疗提供了坚实的技术支持。

综上所述，基因组学技术的发展涵盖了从测序技术到非测序技术的多个层面，为染色体病的研究提供了丰富且多样化的工具支持。这些技术的协同应用与不断优化，不仅显著提高了染色体病的诊断精度，还为其分子机制的深入研究开辟了新的路径，推动了该领域的发展。

2、简述单基因病临床发病特点（和其他类型遗传病相比较）。

遗传病是由遗传物质异常改变（染色体畸变和基因突变）所致的疾病，可分为单基因遗传病、多基因遗传病和染色体病。如果一种遗传病的发病仅仅涉及到一个基因，这个基因称为主基因，其导致的疾病称为单基因遗传病。单基因病的临床发病特点与染色体病和多基因遗传病相比具有一定的独特性，主要包括以下几个方面：

① 遗传率高，呈现家族聚集性：单基因病通常遵循经典的孟德尔遗传规律，其遗传方式明确，可分为常染色体显性遗传、常染色体隐性遗传、X 连锁显性遗传、X 连锁隐性遗传以及线粒体遗传。在这种疾病中，遗传基础的作用占主导地位，遗传率接近 100%，环境因素的影响相对较小。由于遗传方式多样，因此在外显率方面，单基因病既可能表现为完全外显，如家族性腺瘤性息肉病，也可能为不完全外显，表现出携带突变的个体中部分未显现相关症状的情况，例如多囊肾病（PKD）。相比之下，染色体病的外显率通常接近 100%，尤其是整条染色体的数目异常，如 21 三体综合征。但部分嵌合型染色体病患者可能表现出部分外显；而多基因病的外显率通常较低。

由于单基因病遵循明确的遗传模式，患者家族中通常会有类似病例或相关携带者，具有较强的家族聚集性。如在近亲婚配的家系中，常染色体隐性遗传疾病的发生率就会显著升高。因此，家族系谱分析能帮助区分单基因病与多基因病。

② 致病基因明确，表现度变异范围小，存在遗传早现：单基因病由单一基因的功能丧失或获得性功能突变引起，从而导致基因产物（如蛋白质或 RNA）的异常表达或功能紊乱。单基因突变所引发的症状通常较为严重，且表型直接与突变的基因功能缺陷密切相关，表现为特定的临床症候群，显示出显著的特异性。如马凡综合征患者（第 15 对染色体上的纤维蛋白 FBN1 发生变异）常表现为高瘦的身材、四肢修长以及心血管异常。这种与致病基因功能密切相关的典型表型为疾病的临床诊断提供了重要依据。由于具体表型常与单一基因的功能缺陷直接相关，因此表现度变异范围相对较小。与此相比，染色体病表现度变异性通常更大，尤其是部分性染色体异常或嵌合型染色体病患者；而多基因病因由多个微效易感基因和环境因素协同作用导致，因此症状严重程度和发病

模式因遗传背景和环境要素的不同呈现显著不同。

此外，遗传早现是单基因病中的一个显著特点，尤其是在三核苷酸重复扩增病中，如亨廷顿舞蹈病和肌强直性营养不良，突变的三核苷酸重复数会逐代增加，从而导致发病年龄提前或症状加重。相较之下，染色体病很少出现遗传早现，因为其病因通常是随机发生的染色体异常，而非突变累积或扩增的结果；多基因病在遗传早现方面也不常见，尽管某些疾病如癌症风险可能在家系中逐代升高，但这种趋势更多与环境暴露累积或筛查习惯的变化相关，而非遗传意义上的早现。

③ 发病年龄集中：染色体病的发病通常集中在胚胎发育早期至出生阶段，因其导致严重的染色体结构或数目异常，常引发流产、死胎或先天性畸形等严重后果。一些存活的新生儿可能表现出明显的染色体异常综合征的特征，如唐氏综合征，其临床症状在出生后即可被明确识别；而多基因病的发病机制更为复杂，因涉及多个基因与环境因素的交互作用，其风险因素在个体成长过程中逐渐累积，往往在成年后才开始显现症状，例如高血压、糖尿病等。多基因病的病程具有渐进性，其发病时间和临床表现往往存在较大的个体差异。

与上述两种类型的遗传病对比，单基因病的发病时间通常集中于出生后至青春期前的儿童期。这一特点与其致病机制密切相关：单基因病多由单一基因的功能丧失或获得性突变引起，导致关键生理功能的显著异常。儿童期正处于生长发育的关键阶段，许多与发育相关的基因在这一时期的功能尤为重要，因此当关键基因发生突变时，相关疾病在儿童期较易表现出明显的症状。不过，某些单基因病会表现为延迟显性，即患者在早期发育阶段可能无明显异常，直到成年后才逐渐显现症状（如遗传性舞蹈病）。这类疾病的延迟发病特性通常不影响个体生育能力，使得致病突变在家族中具有更高的遗传传递概率，因而家族遗传史在诊断中具有重要参考价值。

④ 发病率及发病人群具有特色：单基因病的发病率相对较低，但每一种单基因病的特定病例在家族中具有明显的遗传性。虽然单基因病的总体发病率低，但由于遗传模式清晰，其在一些高风险家系中的发病概率会显著增加。

与此对比，染色体病的发病率总体同样较低，在高龄孕妇中较为高发。随着母亲年龄的增加，卵母细胞减数分裂过程中染色体分离错误的概率随年龄增

长而增加，从而提高染色体数目或结构异常概率，导致胚胎早期流产。因此发病率增高；而多基因病的发病率相对较高，在人群中分布广泛，且发病率因种族、地区和生活方式的的不同而有显著差异。

3、简述一代、二代、三代测序技术原理及用于分子诊断的选择依据。

① 一代测序（Sanger 测序）：Sanger 测序的核心原理是通过引入双脱氧核苷酸（ddNTPs）终止 DNA 链的合成，从而获得一系列长度不等的 DNA 片段，并利用荧光标记和电泳技术解析 DNA 序列。

具体而言，测序反应以单链 DNA 为模板，在 DNA 聚合酶的作用下，以引物和四种单脱氧核苷酸（dNTPs）作为底物进行链的延伸。此时在反应体系中加入适量的荧光标记双脱氧核苷酸（ddNTPs），这些分子因缺乏 3'-羟基，会在随机掺入 DNA 链时终止链的延伸。由于每种 ddNTP 与对应的 dNTP 共同存在于反应体系中，链终止便会发生在不同的位点，从而生成以共同引物为 5'端、以随机掺入的 ddNTP 为 3'端的一系列长度不等的 DNA 片段。这些片段随后可通过毛细管电泳分离，根据片段的长度差异（相邻片段仅相差一个碱基）进行排序。由于每种 ddNTP 均携带特定的荧光标记，通过在电泳过程中通过荧光探测器检测其末端荧光信号，我们便可以识别片段终止时碱基的种类。

通过依次读取这些片段的终止碱基，测序仪器可以准确重建目标 DNA 序列的碱基排列顺序。

② 二代测序：二代测序以 Illumina 测序平台为代表，其核心原理是桥式扩增和边合成边测序。测序过程包括文库制备、DNA 簇生成和测序反应三个主要步骤。通过结合高效扩增和精确信号，二代测序能够实现大规模、高通量的 DNA 检测。

(1) 文库制备：测序的第一步是对基因组 DNA 进行片段化处理。将目的 DNA 打断为小片段，随后进行末端修饰，即在 5'端加上磷酸基团、3'端加上腺嘌呤以补平末端，并在两端连接特定的 DNA 接头序列，接头序列与测序平台流动槽（Flow Cell）表面固定的引物序列互补。文库制备的目的是确保每个 DNA 片段都能与测序系统配套，同时便于后续的固定和扩增。

(2) DNA 簇生成：文库中的 DNA 片段被固定到 Flow Cell 表面——Flow Cell 是带有微通道的厚玻璃片，表面共价连接了两种不同的引物序列。待测 DNA 片段通过两端接头序列，与固定在 Flow Cell 表面的引物序列进行杂交。在 DNA 聚合酶的作用下，以原始 DNA 片段为模板，合成第一条互补链。延伸

完成后，原始 DNA 链被洗去，仅保留新合成的单链 DNA 分子。随后，固定在 Flow Cell 表面的单链 DNA 分子弯曲并杂交到邻近位置的引物上，形成类似“桥”的结构。在 DNA 聚合酶的催化下，杂交后的引物以单链 DNA 为模板完成延伸，形成双链 DNA 桥式结构。通过变性作用，双链 DNA 桥分离成两条单链，其中一条单链被剪切并洗去，仅保留另一条单链，同时 3'端被化学封闭以防止非特异性延伸。这一桥式扩增过程反复进行，循环的变性、杂交和 DNA 合成反应使得待测 DNA 片段逐步扩增，最终在 Flow Cell 的每个位置生成由大量相同 DNA 分子组成的 DNA 簇。这些 DNA 簇显著增强了荧光信号的强度，确保测序系统能够准确检测每个碱基的信号变化，为后续的高通量测序提供可靠的模板支持。

(3) 测序反应：Illumina 测序技术采用了基于可逆终止的边合成边测序（SBS）技术，通过使用荧光标记的可逆终止核苷酸，实现单碱基的逐步合成与实时检测。在每轮合成反应中，每种 dNTP 都携带特定颜色的荧光标记，其 3'端被可逆化学封闭，防止多次碱基掺入，从而确保每次反应仅能合成一个碱基。在 DNA 聚合酶的作用下，新加入的 dNTP 与模板链配对后，通过荧光成像仪检测每个 DNA 簇的位置及其发出的荧光信号颜色，从而确定所掺入碱基的种类。完成荧光检测后，通过化学处理切除 dNTP 上的荧光标记和 3'端的阻断基团，暴露新的 3'-OH，以便启动下一轮的碱基合成。

因此，测序循环由三个主要步骤组成：dNTP 掺入、荧光信号成像和阻断基团去除。通过 SBS 技术，Illumina 平台能够同时对数百万至数十亿条 DNA 链进行测序，为大规模基因组分析提供了强有力的技术支持。

对于双端测序，DNA 片段的两个末端分别被读取。在完成第一轮测序（Read 1）后，生成互补链并清除已测序链，然后继续合成另一条链（Read 2），确保两端序列的读取完整性。

综上所述，Illumina 测序平台通过桥式扩增生成高密度 DNA 簇，实现了 DNA 片段的大规模扩增；通过边合成边测序原理，精确检测每个碱基的荧光信号。此技术不仅具有高通量、高准确度的特点，还能通过双端测序获得更全面的序列信息，是现代基因组学研究的重要工具。其余的二代测序平台如 Roche 454、Ion Torrent 大多也是基于边合成边测序实现，故在此不赘述。

③ 三代测序：三代测序技术的主要方法包括单分子实时测序（SMRT）和纳米孔测序，两者在测序原理和技术实现上各具特色。

SMRT 的核心原理是基于零模波导（ZMW）技术实现的单分子测序。具体而言，待测 DNA 片段通过加接头形成环形 DNA 模板（SMRTbell），在 ZMW 反应室内以单分子形式结合到固定的 DNA 聚合酶上。测序反应中使用带有荧光标记的 dNTPs，DNA 聚合酶将 dNTPs 逐一掺入新合成的链中。当一个碱基掺入时，荧光信号被实时激发并通过光学检测器捕获，随后荧光标记被快速剪切并释放，以便进行下一轮反应。由于测序过程是连续且实时的，SMRT 能够直接测定 DNA 的碱基序列，同时还可以通过测量 DNA 聚合酶的停顿时间检测表观遗传修饰（如甲基化）。这种技术的显著优势在于其读长极长（通常超过 10 kb，甚至达到 100 kb），能够高效解析复杂基因组结构（如重复序列和大片段结构变异），但其准确率较低，需通过重复覆盖和数据纠错提高精度。

纳米孔测序的原理是基于电信号的实时监测：待测 DNA 分子首先通过预处理形成适合测序的片段，随后在测序装置中通过纳米孔进行测序。纳米孔是一种嵌入膜中的蛋白质通道，测序时通过施加电场，可将单链 DNA 分子拉入纳米孔。在 DNA 分子通过孔洞的过程中，不同碱基会对离子流产生特定的阻碍作用，从而会引起电信号的变化。这些变化由传感器实时记录，最后通过算法解析出原始的 DNA 序列。与 SMRT 技术不同，纳米孔测序无需荧光标记或酶促反应，因此设备简单，测序速度快。同时，纳米孔测序还能够直接读取 RNA 序列及其修饰。然而，由于电信号解读的复杂性，纳米孔测序在准确性上仍面临一定的挑战。

总之，三代测序技术通过单分子实时测序和纳米孔测序的创新，实现了对长读长、高重复序列和表观遗传信息的解析。这些技术显著提升了基因组组装和结构变异检测的能力，为复杂基因组研究和个性化医学提供了强有力的技术支持。

④ 分子诊断时使用测序方法的依据：基因组测序技术在分子诊断中的选择主要取决于检测目标、分辨率需求、通量、成本以及数据分析的复杂性。一代测序、二代测序和三代测序各有其独特优势和适用场景：

一代测序作为基因组学研究的奠基技术，以其高准确性成为单基因突变检

测的金标准。由于其对单碱基替换、小片段插入或缺失具有高度敏感性，一代测序适用于已知基因突变位点的检测，尤其在单基因遗传病、靶向治疗中的基因靶点检测（如 **EGFR** 突变）中被广泛使用。此外，一代测序常用于验证二代或三代测序发现的变异，确保结果的可靠性。然而，其低通量和高成本限制了在全基因组层面的大规模应用，因此更适合明确检测目标的小范围精准分析。

相比之下，二代测序（**NGS**）凭借高通量和高覆盖率的特点，显著扩展了分子诊断的范围。通过快速测序全基因组，**NGS** 能够高效识别点突变、小片段插入缺失以及低频变异，在肿瘤分子诊断、多基因病的基因 **panel** 检测以及外显子组测序（**WES**）中表现出强大优势。此外，**NGS** 可以通过算法推断大规模染色体变异（如倒位、重排和拷贝数变异），尽管这一过程依赖计算分析而非直接观察，这使其在处理复杂结构变异或高重复序列时存在一定局限性。总之，二代测序适合需要大规模、多基因检测的场景，但较高的成本和复杂的数据分析对实验室的技术能力提出了更高要求。

三代测序技术通过其长读长特点弥补了二代测序的不足，成为复杂结构变异和大规模染色体变异研究的重要工具。单分子实时测序（**SMRT**）和纳米孔测序技术能够直接解析基因组中的大片段重排、倒位以及高重复序列，尤其在解析复杂基因组或疑难遗传病时表现出显著优势。此外，三代测序能够直接检测表观遗传修饰（如 **DNA** 甲基化），拓展了分子诊断的深度。然而，三代测序成本较高，准确率仍有待提升，数据分析的复杂性也限制了其广泛应用，通常用于高度复杂或难以解决的诊断需求。

在分子诊断中，这三类测序技术各自适用于不同场景。一代测序因其高精度适用于小规模、单基因检测；二代测序以其通量高、范围广成为多基因检测和基因组水平分析的主力；三代测序则凭借长读长和表观遗传检测能力，在复杂基因组结构解析和疑难病例中占据优势。选择适宜的测序技术需综合考虑检测目标的规模与复杂性、成本、时间以及实验室能力，确保在分子诊断中实现最优的检测效果。这些测序技术的互补性也为精准医学的发展提供了重要支持。

4、请简述人类参考基因组的研究方法和历程，并结合目前最新的研究进展（如 T2T 和泛基因组），举例说明其在神经发育障碍等复杂疾病遗传学研究中的应用。

人类参考基因组的研究方法主要包括两种：人类基因组计划（HGP）及 Celera 公司采用的基因组图谱绘制方法。两者在技术路线和实施细节上有所不同，但都可构建出高质量的人类参考基因组。

① HGP 是 20 世纪 90 年代初启动的一项国际科研合作项目，目标是绘制出完整的人类基因组图谱。该计划采取了以下几种主要方法实现目标：

(1) 构建遗传图谱：在真核生物的减数分裂过程中，染色体会发生重组，重组的概率随两基因之间距离的增加而变化。基于这种重组率，可以计算基因之间的遗传距离（通常用 cM 表示，1 cM 对应 1% 的重组率）。遗传图谱的构建基于上述遗传学原理，通过遗传标签（例如特定基因的标记或 SNP）来确定基因在染色体上的相对位置，从而绘制出基因间的连锁关系。通过测定不同基因之间的重组率，科学家能够绘制出基因在染色体上的相对位置图。这种图谱只能显示基因之间的相对距离，但不能揭示精确的物理位置。

1994 年 9 月，HGP 团队完成了基于 3000 个遗传标记，分辨率为 1 cM 的遗传图谱绘制，为后续基因组研究奠定了基础。

(2) 构建物理图谱：物理图谱通过已知的 DNA 序列标记（如限制性酶切位点、序列标签位点等）来绘制基因组图谱，图距单位通常为 Mb（百万碱基对）或 Kb（千碱基对）。构建物理图谱的方法主要包括限制性酶切作图、克隆作图、荧光原位杂交（FISH）作图和 STS（序列标记位点）作图。HGP 团队主要采用 STS 作图方法。STS 是一种长度为 200-500bp 的短 DNA 序列，通常在基因组中只有一个拷贝。通过 PCR 技术和分子杂交方法，STS 标记可以精确地定位基因组中的特定 DNA 序列。

与遗传图谱通过重组率来计算基因间距离不同，物理图谱是通过随机切割 DNA 并计算两个标记位点所在片段的重叠概率来确定基因之间的物理距离。两个标记位点距离越近，它们出现在同一片段的概率越高。基因组物理图谱与遗传图谱结合，能够更准确地确定基因之间的位置关系。

1998年10月，人类基因组计划完成了包含 52,000 个序列标签位点的物理图谱的绘制，为基因组测序提供了重要的指导。

(3) 测序鉴定：在 HGP 中，测序与鉴定是确保基因组被精确解读的核心步骤。基于前期构建的遗传图谱和物理图谱，HGP 团队采用了“重叠群克隆”技术，并结合亚克隆测序、组装与校对，最终获得了高精度的人类基因组序列。

首先，HGP 使用了重叠群克隆技术，通过将基因组 DNA 分割成多个较大片段，并插入到细菌人工染色体（BAC）和酵母人工染色体（YAC）克隆载体中，对细胞进行扩增和储存。BAC 和 YAC 分别能容纳 100-300kb 和 600kb-1Mb 的 DNA 片段，使得每个克隆能够覆盖基因组中不同的区域，并且每个克隆之间具有重叠部分，确保后续组装时能够精确拼接。

接下来，HGP 团队进行亚克隆测序。具体而言，首先通过酶切或其他方法将这些大克隆拆分为多个较小的 DNA 亚克隆，并对这些小片段进行 Sanger 测序。每个小片段的序列信息通过比对重叠区域进行拼接，基于其在染色体上的物理位置，将拼接后的序列与已知的遗传图谱进行对照校正，从而形成完整的基因组序列。

2003 年 4 月，HGP 成功完成了人类基因组 98% 的序列测定，精度高达 99.99%。尽管仍有少数高变异区域未被完全测定，但该项目的成果为后续的基因组学研究、个体化医疗及疾病研究提供了重要的数据支持，极大地推动了生命科学和医学的发展。

② Celera 公司采用了一种不同于人类基因组计划的方法——全基因组鸟枪法。该方法通过将整个基因组随机打碎成大量的小片段进行测序，然后通过计算机算法将这些短片段进行拼接与重组，从而确定它们在基因组中的准确位置。全基因组鸟枪法的最大优势在于它不依赖于事先构建遗传图谱或物理图谱，因此具有更高的灵活性。这使得它成为一种快速有效的基因组组装方法，尤其适用于基因组结构复杂或未知的物种。该方法的核心是基于重叠序列的拼接，将大量短序列拼接成较长的连续序列（contig）。通过这些 contig 的拼接，科学家能够推断出整个基因组的序列。

虽然全基因组鸟枪法相比人类基因组计划的方法具有更少的前期准备，但其拼接过程的复杂性要求强大的计算能力。Celera 公司采用了先进的计算技术，

成功地实现了基因组的组装。

上述两个 2003 年发表的人类全基因组图谱都只覆盖了基因组中常染色质的部分，而异染色质区域的序列探究尚未完成。为完成基因组最后剩余区域的组装，2022 年研究者发布了“端粒到端粒”（T2T）的人类基因组图谱，填补以往基因组测序中未解析的复杂区域，尤其是端粒、着丝粒和高重复序列区域。T2T 图谱采用纯合 CHM13hTERT 细胞系进行测序，通过 PacBio HiFi 长读长测序和 Oxford Nanopore 超长读长测序获得高精度、长读长的序列信息；结合 Illumina 短读长测序提供短读长的高覆盖数据、Arima Genomics Hi-C 三维基因组测序检测基因组三维结构以及单细胞 DNA 模版链测序辅助鉴定细胞的染色体结构和变异，并构建了识别重复序列的方法，完成了对人类 rDNA 阵列及其周围序列的组装，从而完整地覆盖并解析包括端粒和着丝粒在内的所有人类基因组区域，克服了传统测序方法无法解析的重复和复杂区域。

T2T 基因组在神经发育障碍研究中的作用具有重要的意义，特别是在探索基因组的复杂性和遗传架构方面。以常见的神经发育障碍——自闭症谱系障碍（ASD）为例，ASD 的高度遗传性受常见和罕见变异的共同影响。先前研究发现，尽管这些常见和罕见的编码区变异对 ASD 的发生有一定影响，但它们不能完全解释 ASD 的高度遗传性，仍然存在大量的遗传因素未被识别。罕见的非编码变异可能在 ASD 的遗传架构中发挥重要作用。T2T 基因组通过提供更加完整和精确的基因组序列，能够解析以往参考基因组（如 GRCh38）未能涵盖的重复区域和结构复杂区域，更全面地分析包含非编码区域的变异在内的难以解析的区域，从而进一步深入揭示影响 ASD 风险的潜在遗传因素。

在人类基因组中，存在大量的结构变异（SVs），这些变异在不同人群中表现为多态性。相比于单一的 SNPs 或 indels，拷贝数变异（CNVs）和基因组结构变异（gSVs），更有可能影响基因的功能。传统的参考基因组，包括 GRCh38 和 T2T 基因组，并不能涵盖所有的结构变异，导致超过三分之二的结构变异被遗漏。这些未被识别的变异使得基因组中重要的遗传信息无法被准确捕捉。因此，传统参考基因组在复杂的重复区域和大规模结构变异的分析中存在显著局限。为解决这一问题，泛基因组学应运而生。泛基因组学旨在通过综合多个个体的基因组数据，构建一个包含全球不同种群遗传多样性的参考基

基因组。2023 年，人类泛基因组参考联盟（HPRC）发布了泛基因组的首个草图版本，构建了一个更加全面的基因组参考框架。这一框架不仅能够更好地捕捉全球范围内的遗传多样性，还能有效提高基因组分析的精度，弥补传统参考基因组的不足。

在神经发育障碍疾病中，CNV 和 gSV 均扮演着至关重要的角色。研究显示，罕见 CNV 在所有 ASD 病例中占 5-10%，负担显著高于对照组——即 ASD 患病者往往比其未患病的兄弟姐妹携带更多的罕见 CNV，并且通常影响一个或多个基因。而新生 CNV（即父母未携带的变异）则与 ASD 的风险密切相关。此外，gSV——尤其是新生 gSV，也在 ASD 患者中呈现出较高的负担，并且通过直接破坏基因或影响基因表达来对基因功能产生重大影响。尽管 gSV 的规模较 CNV 小，但它们在人群中更为普遍，对基因组的功能有着深远的影响。由于传统的参考基因组无法完整覆盖这些结构变异，从而限制了对 ASD 的深入研究。而通过泛基因组学方法，研究者能够获得包括结构变异在内的更加全面的遗传信息，识别与这些遗传变异相关的表型变化，突破现有参考基因组的局限，助力基因定位及功能位点的挖掘，为理解 ASD 的遗传机制提供新的视角。

此外，不同人群在基因组上存在显著的遗传差异，特别是结构变异（SVs）的多样性，这对于 ASD 等神经发育障碍的研究至关重要。目前，许多人类基因组学研究和公共数据库的数据大部分来源于欧美人群，非欧美人群的遗传数据较为匮乏。通过泛基因组研究，研究者或许可以发现中国人群基因组中与神经发育障碍特有的 SV 热点，从而为中国人群的精准医学和个体化治疗发展奠定基础。

综上所述，T2T 基因组和泛基因组学有望成为探索 ASD 遗传机制和解决 ASD 高遗传度问题的重要工具。

5、简述全基因组关联分析算法的基本流程。

GWAS（全基因组关联研究）的基本流程可以分为以下几个步骤：

构建队列、收集表型信息：全基因组关联研究（**GWAS**）的第一步是确定研究对象，并据此选择适当的表型变量。表型通常分为二分类变量和连续型变量。二分类变量指的是具有两个可能结果的变量，如性别、疾病或死亡等；而连续型变量则是可以在某一范围内取任意值的变量，如身高、体重、年龄等。

在 **GWAS** 中，选择合适的研究人群至关重要。因为该研究通常需要较大的样本量以确保显著关联的可重复性，研究者可以使用功效计算软件（如 **CaTS** 或 **GPC**）来确定所需样本量。根据研究目标，研究可采用前瞻性或回顾性队列设计。前瞻性队列研究通常根据研究对象的暴露情况进行，适合研究尚未发生的结局，但需要较长时间和较大样本；回顾性队列研究则是基于已有的表型或疾病进行，能快速获取结果，但可能存在资料控制不足的偏倚。确定队列后，便可以从队列个体中收集 **DNA** 样本，并记录相应的表型信息。

目前许多 **GWAS** 都利用现有公共资源数据集进行分析，需要注意在数据收集和招募过程中避免潜在偏倚，充分考虑种群结构，以避免假阳性结果。

基因分型：在 **GWAS** 中，个体的基因分型可通过微阵列或 **NGS** 技术实现。微阵列技术这一方法主要用于检测常见变异。由于微阵列技术只能覆盖有限的 **SNP** 位点，因此需要进行基因型填补，以填补芯片未能覆盖的 **SNP** 位点。基因型填补通过单倍型推断技术，精确预测未被芯片设计覆盖的 **SNP** 位点的基因型，从而扩大可用于关联分析的遗传位点范围，提高发现潜在致病基因的可能性。下一代测序（**NGS**）技术，如全外显子组测序（**WES**）和全基因组测序（**WGS**），能够提供更全面的基因组信息，涵盖罕见变异，并且具备较高的通量，能够同时进行 **SNP** 及基因组结构的注释。然而，由于 **NGS** 的成本较高，微阵列基因分型仍是 **GWAS** 中最常用的分型方法。

质量控制：**GWAS** 的输入文件通常包括匿名的个体 **ID**、个体之间的家庭关系编码、性别、表型信息、协变量、所有变异的基因型数据以及基因分型批次信息。在数据输入后，为了获得可靠的 **GWAS** 结果，必须进行严谨的质量控制。质量控制步骤包括去除稀有或单一等位基因的变异、去除不符合哈迪-温伯格平衡的变异、过滤掉在部分个体中缺失的 **SNP**、识别和去除基因分型错误，并确

保表型数据与基因型数据的匹配——通常需要通过比较自报性别与基于 X 和 Y 染色体推测的性别来校验。为了执行这些质量控制步骤，在 GWAS 的质量控制中，多个软件工具可用于质控过程：**PLINK/PLINK2** 可用于筛选不合格的 SNP 和个体，检查基因型的哈迪-温伯格平衡偏离、基因分型调用率等，同时进行性别校验和亲缘关系检查；**RICOPILI** 主要用于对原始基因数据及元分析中的汇总统计数据进行了质控，确保输入数据的可靠性；**SMARTPCA** 通过主成分分析（PCA）来校正群体分层的影响，帮助消除由群体结构引起的假阳性结果；**FlashPCA** 与 **SMARTPCA** 类似，但在处理大规模样本时具有更高的速度和可扩展性，适用于大样本数据集的主成分分析。这些工具在 GWAS 的质控过程中发挥着至关重要的作用。

群体结构校正：在 GWAS 中，群体的祖先和亲缘关系需要谨慎考虑，特别是在来自不同背景的参与者数据集中。未能考虑群体分层可能导致假阳性或假阴性遗传信号的出现，以及偏倚的统计检验结果，从而影响遗传信号的解释。GWAS 通常通过主成分分析（PCA）来考虑祖先影响。在此过程中，首先利用所有个体的基因型数据进行聚类分析，以识别和排除离群值，然后将主成分作为协变量纳入后续的 GWAS 回归模型，从而校正群体结构对分析结果的影响。这一过程有助于确保 GWAS 结果的准确性和可靠性。

基因型补充：该步骤用于推测未基因分型的单核苷酸多态性（SNP）的基因型。补充过程首先通过选择适当的参考人群面板和目标人群基因型数据进行比对，使用如 **IMPUTE2**、**BEAGLE**、**MACH** 和 **SHAPEIT2** 等工具进行补充。补充步骤包括几个关键环节：首先，需要对个体基因型进行统计学计算，确定基因型的推测值（单倍型分相）；然后，根据经过决定是使用硬性基因型调用还是加权调用；随后，选择适当的参考人群面板，并解决可能出现的链向问题及不同平台间的差异，必要时移除模糊的 SNP，并检查小等位基因频率及连锁不平衡模式，确保参考面板与目标数据的一致性。最后，进行缺失基因型的补充，并对补充质量进行评估。基因型补充能够提高 GWAS 数据的覆盖度，增强研究结果的精确性和可靠性。

遗传关联分析：遗传关联分析通过回归模型检验基因型与表型之间的关联。连续性表型使用线性回归，而二元表型使用逻辑回归模型进行检验。此处的

GWAS 线性回归模型为：

$$Y \sim W\alpha + X_s\beta_s + g + e \quad (1); \quad g \sim N(0, \sigma_A^2 \psi) \quad (2); \quad e \sim N(0, \sigma_e^2 I) \quad (3)$$

其中， Y 是表型值的向量， W 是包含截距项的协变量矩阵， α 是协变量的效应大小向量， X_s 是个体在 SNP 位点 s 的基因型值向量， β_s 是遗传变异 s 的效应大小（即 SNP 效应大小）， g 是随机效应，表示其他 SNP 的多基因效应， e 是残差误差， σ_A^2 表示表型的加性遗传变异， ψ 是标准遗传关系矩阵， σ_e^2 表示残差方差， I 是单位矩阵；在逻辑回归模型中，使用 **logit** 连接函数来建模二元分布的病例对照表型的结果概率。

为提高统计功效并减少群体分层的影响，可以在混合模型中加入随机效应项，以捕捉个体间的遗传相关性（使用 **fastGWA** 实现）。

假阳性控制与效果估计校准：该步目的是减少在大规模基因关联研究中由于多重假设检验带来的假阳性结果。由于 **GWAS** 通常涉及数百万个基因变异与表型之间的关联检验，因此需要设定严格的多重检验阈值。常用的 **Bonferroni** 修正阈值为 $P < 5 \times 10^{-8}$ ，以控制假阳性率。这个阈值可能会根据人群特征有所不同，特别是在较大有效种群或较小等位基因频率的情况中，需调整阈值以减少多重检验负荷。

对于多基因遗传性表型（如身高、精神分裂症、2 型糖尿病），往往涉及许多效应较小的基因变异，因此会出现赢家的诅咒现象，即在初步 **GWAS** 中，接近发现阈值的效应值通常被高估。为了控制假阳性和赢家的诅咒，需要比较发现队列和独立验证队列之间的效应大小。验证队列应在 **GWAS** 开始时就考虑，并具有足够的统计功效进行校准。

统计精细定位：精细定位是通过分析连锁不平衡和关联信号，结合基因功能注释等信息，确定最具可能性的因果致病变异，为进一步的生物学研究提供关键线索。在 **GWAS** 中，由于连锁不平衡，许多非致病变异可能与表型显著相关，因此，需要进一步分析以确定哪些变异可能是因果变异。

精细定位分析的常见方法包括条件关联分析，旨在通过在基因型-表型回归模型中加入主导变异作为协变量，调整区域内的关联信号。当多个关联信号存在时，通常采用逐步选择法，直到没有显著关联剩余。除了条件关联分析，还有基于贝叶斯模型的精细定位方法，如 **CAVIAR**、**FINEMAP** 和 **PAINTOR**，这

些方法通过先验概率分布来优化回归模型变量的选择，相较于传统的条件关联分析，贝叶斯模型能考虑更多信息（如估算准确度）。

在精细定位中，重要的是扩大评估变异的覆盖范围，使用全基因组测序（WGS）为基础的基因型填补参考面板——如 gnomAD 和 TOPMed，可以进一步提高精细定位的准确度。

GWAS 的功能推断：此步骤旨在将 GWAS 识别出的统计学关联转化为有意义的生物学信息，具体包括确定与复杂性状相关的基因、调控机制和细胞通路。虽然 GWAS 能够识别出成千上万与特征相关的遗传变异，但这些变异的生物学意义通常难以直接推断。功能推断的目标是识别潜在的致病变异，分析这些变异对基因调控的影响，并理解其在分子和生理层面的下游效应。

在功能推断中，首先需要确定可能受影响的基因，尤其是当变异位于编码区外时。常用的方法包括分子定量性状位点（molQTL）分析、eQTL 分析和增强子-基因关联分析，这些方法可以将变异与基因表达及其他分子表型联系起来。然而，这些方法也存在局限性，尤其是在处理那些不直接影响蛋白质编码区域的变异时，或在特定组织和细胞类型的数据稀缺时。为了提高基因优先级排序的准确性，还需要整合来自不同数据源的信息，如染色质构象捕获技术和实验性扰动。最终，功能推断的目标是识别与目标性状相关的调控通路和细胞效应，全面理解遗传变异如何在分子和细胞层面上影响生物学过程。这一过程对于推动个性化医疗和治疗干预具有重要意义。

最后，全基因组关联分析（GWAS）的结果展示方式主要包括曼哈顿图和 Q-Q 图等，这些图形工具帮助研究者直观理解遗传变异与表型之间的关系。曼哈顿图是 GWAS 中最常用的结果展示工具之一，它通过表示大量非显著数据点（数值较低变化较大）以及少数几个显著数据点的簇，使得显著点在图中呈现为高耸的塔状结构。曼哈顿图主要用于绘制 p 值，通过使用 $-\log_{10}(p \text{ 值})$ 转换，使得较小的 p 值在图中显示得更高，形成显著的“塔楼”状结构。；Q-Q 图主要用于从模型的角度看一下显著位点是否是假阳性。它通过比较观察值和期望值的分布，帮助研究者判断显著性位点的可靠性；LDblock 绘制连锁不平衡和单体型图，有助于理解遗传变异之间的关联性；LD 衰减图用于展示连锁不平衡衰减距离，有助于确定显著 SNP 位点之间的 LD 关系以及候选基因内各 SNP

位点之间的 LD 关系。

全基因组关联分析的荟萃分析（GWAMA）：GWAMA 通过整合多个队列的数据来扩大样本量，从而提高 GWAS 的统计功效。在 GWAMA 中，通过工具如 METAL、N-GWAMA 或 MA-GWAMA 进行数据分析，并使用诸如 RICOPILI 或 EasyQC 等质控管道确保数据质量。GWAMA 中的各队列需遵循统一的数据分析流程，以便不同队列之间的效应大小可以进行比较（这包括将效应大小标准化为标准正态分布，因为不同队列的表型测量和效应大小可能无法直接比较）。再通过固定效应模型或随机效应模型进行荟萃分析，结合各队列的贡献，从而更准确地估计效应大小和统计显著性：固定效应模型假设各队列的误差方差相等，而随机效应模型则用于检验结果是否存在异质性。通过加权合并各队列的结果（通常按样本量或使用逆方差法），GWAMA 可以提供更为精确的效应估计。

GWAS 因能够识别与疾病或性状相关的遗传变异，故被广泛应用于风险预测和遗传结构理解。通过多基因风险评分（PRS），GWAS 结果可用于预测疾病风险，辅助临床筛查，并提高疾病预防效果；除此之外，GWAS 还帮助估计性状的遗传架构，包括致病变异的数量、效应和频率，进而揭示遗传相关性和遗传关联的复杂性。在临床实践中，GWAS 推动了个体化医疗的发展，尤其在评估疾病遗传风险及其预后中发挥重要作用。

6、什么是肿瘤的异质性？导致肿瘤发生的内因和外因有哪些？

① 肿瘤异质性：肿瘤异质性是指肿瘤在发生和演进过程中，由于分子生物学或遗传学层面的改变，不同肿瘤细胞在形态、功能及行为特征上表现出显著差异。这一特性贯穿于肿瘤发展的整个过程，是肿瘤赖以生存、适应和不断演进的核心特性之一。

肿瘤异质性体现在多个层面，包括肿瘤细胞的形态学特征、基因表达模式、代谢特性、运动能力、增殖潜能、侵袭能力以及对药物敏感性的差异。正是由于这些特征的差异，导致不同肿瘤细胞在生长速度、转移能力以及治疗反应方面存在显著差异。根据异质性的发生范围，肿瘤异质性可以进一步分为肿瘤间异质性和肿瘤内异质性。肿瘤间异质性是指不同患者间的肿瘤，或同一患者不同部位的肿瘤在基因组、转录组、蛋白组及表型特征上的差异；而肿瘤内异质性则是指同一肿瘤内部，不同细胞群体在遗传、表观遗传及功能特征上的差异。

肿瘤异质性的形成由多种机制驱动，包括基因突变、基因组不稳定性、表观遗传调控的改变、肿瘤微环境的选择性压力以及肿瘤干细胞的分化等。在这些因素的作用下，肿瘤经历克隆扩增、分支演化以及适应性选择，从而形成高度复杂的异质性特征。基于这些演化过程，可以从三个维度解析肿瘤异质性：空间异质性、时间异质性以及肿瘤细胞与其微环境之间的交互异质性。

空间异质性指肿瘤不同区域内细胞在基因组成、分子特征和功能上的差异。如原发肿瘤的不同部位或原发灶与转移灶之间可能存在显著的细胞亚群分布差异，这些差异可能反映肿瘤在不同局部环境中的适应能力；时间异质性是指肿瘤在演进过程中，基因组及表型特征随时间的累积变化。如早期肿瘤细胞可能因营养充足和生长空间较大而表现为高增殖特性，而晚期肿瘤细胞由于经历了缺氧、营养匮乏及药物治疗的选择压力，可能演化为更加侵袭性强且耐药的亚群。此外，肿瘤细胞与微环境的交互作用也对异质性形成具有重要作用。如肿瘤区域内氧分布的不均可能引起代谢特性的区域性差异，低氧区域的肿瘤细胞通常会激活特定的信号通路，重新编程代谢模式以适应环境，并可能通过免疫逃逸机制进一步增强生存能力。而肿瘤微环境中血管分布不均以及免疫细胞浸润程度的差异，也会加剧肿瘤异质性。

肿瘤异质性的复杂性对癌症的诊断和治疗提出了巨大挑战。由于不同亚群

的肿瘤细胞对治疗的敏感性存在显著差异，某些细胞可能对化疗或靶向药物敏感，而另一些细胞则可能具有耐药性。这种治疗反应的异质性是导致治疗失败及肿瘤复发的重要原因。因此，实现对肿瘤异质性的全面评估与精准刻画，是提高癌症治疗效果、制定个性化治疗方案以及探索肿瘤发生发展机制的关键。

② 肿瘤发生的内因和外因：肿瘤的发生是由内因和外因共同作用的结果。

1. 内因是导致癌变的机体内部各类生物学变化，主要包括体细胞突变、基因组不稳定性、肿瘤干细胞增殖以及组织内间质与实质之间异常的相互作用等多个方面。

体细胞突变是肿瘤内因中的基础因素，主要表现为基因突变和染色体结构的改变。这些突变可分为抑癌基因突变、原癌基因突变及全基因组加倍（WGD）三类，每类突变都在肿瘤发生中发挥着关键作用。

(1) 抑癌基因在维持细胞正常生长和增殖、抑制肿瘤形成方面具有重要功能。当抑癌基因发生功能性突变或损伤时，细胞的生长调控机制可能被破坏，导致细胞无限制增殖并最终形成肿瘤。抑癌基因的功能丧失通常需要“双重打击”机制，即同一基因的两条等位基因均发生突变。这一过程可以通过包括点突变（如无义突变和错义突变）、插入缺失（Indel）导致的氨基酸缺失或移码突变、拷贝数变异（CNV）、获得性单亲二体（通过有丝分裂错误、体细胞重组或双链断裂修复错误等机制实现），以及表观遗传变异（如启动子区域的 DNA 甲基化）等在内的多种途径实现。这些突变机制共同促使抑癌基因完全失活，解除对细胞增殖的抑制，推动肿瘤发生。

(2) 原癌基因的突变通常通过两种方式促进癌变：一是因基因重复或染色体重排导致基因表达水平的上升，该基因编码的蛋白质在细胞中异常积累，从而促进细胞增殖；二是编码蛋白质功能的基因发生突变，导致其功能异常，增强细胞的增殖、分化和存活能力。原癌基因转化为癌基因的突变通常属于获得性显性突变，故即使仅有一个等位基因发生突变，也足以推动正常细胞向癌细胞转化。

(3) WGD 是细胞中整套染色体的复制事件，形成原因主要包括细胞分裂过程中染色体分离错误、复制应激、细胞周期调控失常以及有丝分裂过程中纺锤体组装缺陷等。这些因素可能导致染色体复制不完全或分离不均，从而引发染

染色体组数的倍增。**WGD** 不仅增强了肿瘤细胞的增殖能力，还促进了肿瘤的异质性和适应性，为肿瘤的发展提供了更大的遗传多样性。

基因组不稳定性是肿瘤发生与发展的另一个关键内因之一。体细胞突变事件能够破坏基因组复制与修复的精确调控，导致细胞对基因组稳定性的控制逐渐失效，进而引发更广泛的基因组不稳定性。基因组不稳定性通常表现为多种结构变异，包括染色体断裂、染色体复合体形成以及断裂-融合-桥循环等现象。此外，双链环状染色单体组成的染色体外环状 DNA（**ecDNA**）也是基因组不稳定性的显著标志——由于 **ecDNA** 缺乏着丝粒序列，其在有丝分裂过程中分配不均，导致拷贝数和结构的动态变化。这种变异被认为是肿瘤基因组进化的强大驱动力，其复杂性和拷贝数在肿瘤演化过程中显著增加。

基因组不稳定性在癌症的发生与发展中发挥双重作用。一方面，它加速了 DNA 损伤与突变的积累，诱发更多基因突变；另一方面，这些突变进一步驱动关键癌基因（如抑癌基因和原癌基因）的改变，促进肿瘤的进展。因此，基因组不稳定性不仅是基因突变的结果，也是进一步突变的根本原因。在肿瘤发生与发展的过程中，基因突变与基因组不稳定性形成了因果循环，彼此强化，最终推动肿瘤的持续进化与恶性转化。

肿瘤干细胞在肿瘤的大量扩展中起着关键作用。目前，关于肿瘤干细胞的起源主要存在两种观点。第一种观点认为，干细胞在分裂过程中随机发生的 DNA 复制错误（即随机突变）是肿瘤发生的重要驱动力。根据这一观点，少量因突变而产生的肿瘤干细胞经过演化，逐渐形成具有增殖优势的克隆，从而推动肿瘤的发展。第二种假设则提出，肿瘤内部的异质性细胞群通过表观遗传修饰等机制经历干性重编程，转变为具有癌干性特征的细胞。肿瘤干细胞进一步增加了肿瘤的异质性，增强了肿瘤的长期存活能力和抗治疗能力。

组织内间质与实质细胞之间的异常相互作用也是癌症发生的重要内因。组织间质（如免疫细胞、细胞外基质和成纤维细胞）与实质细胞之间的信号失衡可导致组织结构和形态的异常，从而促进癌症的发生。在这一过程中，细胞增殖被视为一种自然倾向，抑制信号的存在则是控制其增殖的关键。当这些抑制性信号丧失或受到破坏时，细胞增殖失控，最终导致肿瘤的形成。

免疫微环境作为间质的一部分，在肿瘤生成与发展中具有重要作用。正常

情况下，免疫系统通过免疫监视识别并清除异常增殖的细胞，防止癌症的发生。然而，免疫微环境紊乱可能削弱免疫监视功能，导致肿瘤细胞通过免疫逃逸机制规避清除。此外，慢性炎症及免疫细胞（如巨噬细胞和 T 细胞）分泌的促炎因子可破坏间质与实质细胞的相互作用，释放促肿瘤信号，从而推动肿瘤的发生与进展。

2. 肿瘤的发生发展受包括物理性、化学性和生物性致癌因素等多种外因影响。这些外因干扰可以通过不同途径加剧基因突变、基因组不稳定性、肿瘤干细胞的形成和组织微环境的改变，从而推动癌变的发生。

物理性致癌因素，如电离辐射和紫外线辐射，能直接损伤 DNA，增加基因突变和染色体畸变，从而提高肿瘤的发生风险，机械性刺激则通过反复的局部损伤和组织修复过程，促进细胞增殖异常；化学性致癌因素（如重金属、亚硝胺类化合物等）通过与 DNA 结合或代谢成致癌物质，引起基因突变或 DNA 损伤，进而推动肿瘤的形成；生物性致癌因素，特别是病毒、细菌和寄生虫的感染，也是肿瘤发生的重要外因。如人乳头瘤病毒（HPV）可抑制抑癌基因 p53 和 Rb，乙型肝炎病毒（HBV）可通过整合到宿主基因组中促进肿瘤的发生，幽门螺杆菌感染可引起慢性炎症和 DNA 损伤，增加胃癌风险。因此，饮食和生活方式在肿瘤发生中起着关键作用。高脂饮食、过量饮酒和吸烟可增加某些类型癌症的风险，而环境污染——如空气中的 PM2.5 和水中的重金属——也与多种肿瘤的发生密切相关。

7、请举例简述细胞色素 P450 酶系分子基因多态性对药物代谢的影响。

细胞色素 P450 酶系（CYP450）是机体内一类重要的单加氧酶超家族，在生物体的内源性代谢和外源性化合物（包括药物、毒物、环境污染物）的代谢中起着核心作用。这些酶由一组具有遗传相关性的基因编码，按照结构相似性和功能特点被分为多个家族和亚家族，其中 CYP1、CYP2、CYP3 是与药物代谢关系最为密切的基因家族。

药物代谢反应可分为 I 相和 II 相代谢。I 相代谢会添加功能基团或暴露分子中的功能基团，如氧化和还原反应；II 相代谢涉及如葡萄糖醛酸化和磺化等的结合反应。由于 CYP450 的催化功能及其广泛的底物谱，它参与了大多数的药物的 I 相代谢过程。

CYP450 酶系的活性与基因多态性密切相关，其基因编码序列中的遗传变异显著影响药物代谢的速度和效果。这些变异可以导致 CYP450 酶活性的增强、降低甚至丧失，进而形成不同的药物代谢表型。其中，CYP2D6 是这一酶系中的重要成员，其基因多态性直接决定了个体间药物代谢能力的差异，显著影响药物的疗效与安全性。

CYP2D6 的基因多态性包括活性等位基因、部分活性等位基因、失活等位基因和基因复制变异，这些基因型的不同导致了酶活性的显著变化，最终表现为代谢表型的多样性。根据酶活性的高低，个体可分为超快代谢型（UMs）、广泛代谢型（EMs）、中间代谢型（IMs）和低代谢型（PMs）。这些代谢表型的差异直接影响药物代谢速率，进而影响治疗效果与不良反应风险。

首先，CYP2D6 的活性等位基因（如*1、*2、*35）编码具有正常活性的酶，这些等位基因广泛存在于健康人群中，广泛代谢型的个体能够有效地代谢许多常用药物，包括抗抑郁药、抗精神病药和阿片类镇痛药等。例如，在正常代谢能力下，CYP2D6 能将弱效的前体药物（如可待因）转化为其活性代谢产物（如吗啡），从而实现药物的治疗作用。这类个体在药物治疗中表现出较稳定的疗效和较低的不良反应风险；

与正常活性等位基因相比，部分活性等位基因（如*10、*17、*41）则表现为酶活性的降低。这些等位基因在某些人群中具有较高的分布频率，例如*10等位基因在东亚人群中尤为常见。这类基因型导致中间代谢型表型的出现——即这些个体对许多药物的代谢速率显著低于广泛代谢型，但仍具有一定的代谢能力。如在需要服用某些抗抑郁药的情况下，中间代谢型个体服用药物后，相关药物在体内的代谢速率较慢，从而可能引起药物在体内的积累，进而增加不良反应的发生率。在临床实践中，这类个体的药物剂量通常需要调整以避免药物过量或毒性反应；

而失活等位基因（如*3、*4、*5、*6）进一步降低了 CYP2D6 的酶活性。

这些等位基因编码的酶完全丧失功能，导致低代谢型表型的出现。这类个体无法有效代谢许多依赖 **CYP2D6** 代谢的药物，如可待因、**tamoxifen** 等。以可待因为例，低代谢型患者无法将其转化为活性代谢产物吗啡，因此无法获得充分的镇痛效果；而在 **tamoxifen** 治疗中，低代谢型个体无法将前体药物转化为活性代谢产物 **Endoxifen**，从而影响抗肿瘤疗效。对于此类患者，临床上通常建议选择替代治疗方案，例如改用不依赖 **CYP2D6** 代谢的药物，或者直接使用活性代谢产物，以改善治疗效果。

除了活性降低的等位基因外，**CYP2D6** 的基因复制变异（如*1xN、*2xN）会表现为酶活性的增强，形成超快代谢型表型。这类个体因携带多拷贝的活性等位基因，导致 **CYP2D6** 酶活性显著增强，药物代谢速率显著加快。如在使用可待因时，超快代谢型个体能够迅速将可待因转化为高浓度的吗啡，从而导致药物过量相关的不良反应风险，包括严重的呼吸抑制甚至死亡。因此，对于这类患者，药物剂量的个性化调整或选择替代药物尤为关键。

总之，**CYP2D6** 的基因多态性通过改变酶活性直接影响药物代谢能力，其表型差异决定了个体对药物的反应。在药物治疗中，代谢表型的变化可能导致药物疗效显著不同，甚至引发严重的不良反应风险。因此，针对 **CYP2D6** 的基因型检测和药物基因组学指导在个性化治疗中具有重要意义。通过结合患者的基因信息、代谢表型和临床需求，可以更好地优化药物选择和剂量调整，从而提高治疗效果并降低不良反应的发生率。