High-throughput sequencing for biology and medicine

程俊峰

- 高通量测序技术（High-throughput sequencing，HTS）是对传统Sanger测序革命性的改变,一次对几十万到几百万条核酸分子进行序列测定，因此称其为下一代测序技术(next generation sequencing，NGS)，同时高通量测序使得对一个物种的转录组和基因组进行细致全貌的分析成为可能，所以又被称为深度测序(Deep sequencing)。

High-throughput sequencing

Content

- Genomes, variation and epigenomics
- Transcriptomes and other functional elements in genomes
- Medical genomic sequencing
- Single-cell sequencing
- Future developments

- 随着NGS技术的发展，大规模的生物基因组测序变得相当平常。最早是2007年用454技术给一种细菌基因组测序，根据Genomes Online Database的数据，截至2012年6月有3920中细菌和854中真核生物的全基因组被测序。
- 不同的平台有不同的偏好和能力检测变异，但短的基因插入序列（Indels）和较大的结构变异（structural variation，SV）仍然很难侦测；短的读长（reads）可以让我们尝试从头测序DNA（De novo），但是这仍然很具难度，而且拼接时会带来短的叠连群（contigs）。但技术的发展会使测序更精确。

Genomes, variation and epigenomics

- 不断降低的花费和增长的精确度使临床医生和医学研究者可以识别基因易感性标记物和遗传疾病基因特征，辨认受损的编码区或其他功能区的单核苷酸多态性（SNP）是临床基因组学的一部分，研究表明人类个体间大约有3.1-4M的SNPs，有超过30M的SNPs在人类基因组测序项目中被发现：另一个在人类基因组和其他复杂基因组中极具挑战性的领域就是结构变异（SVs），因为SVs在人类健康和疾病调控上有重大的影响。早期的生物芯片技术已经发现人类基因组中存在很多SVs，但NGS技术的发展显示存在的SVs比先前鉴别的更多。

Genome sequence and structural variation

- 新的测序技术使得绘制三维的DNA相互作用图谱成为可能，最早使DNA能在3D尺度上分析的是染色体构象捕捉技术（包括3C、4C、5C，这些技术都不是高通量的）。这些技术的缺点是只能分析已经认为存在作用的位点，而且必须为每个位点设计引物使得通量很低。随着利用高通量的Hi-C技术的产生，更广的染色体基因3D结构图谱绘制成为可能。
- ChIP（chromatin immunoprecipitation，染色质免疫共沉淀）技术和随后的ChIA-PET技术（chromatin interaction analysis by paired-end tag sequencing，配对末端标签测序分析染色质相互作用）使得我们可以研究转录调控因子间的三维的相互作用，比如在染色体上被编码的相距很远的增强子和启动子区域被证明存在很广泛的相互作用。大量的数据分析显示染色体上的不同区域根据生物活性相似性被组织在一起，这些拓扑学的域存在于多细胞型和哺乳类物种。

Mapping higher-order organization in eukaryotic genomes

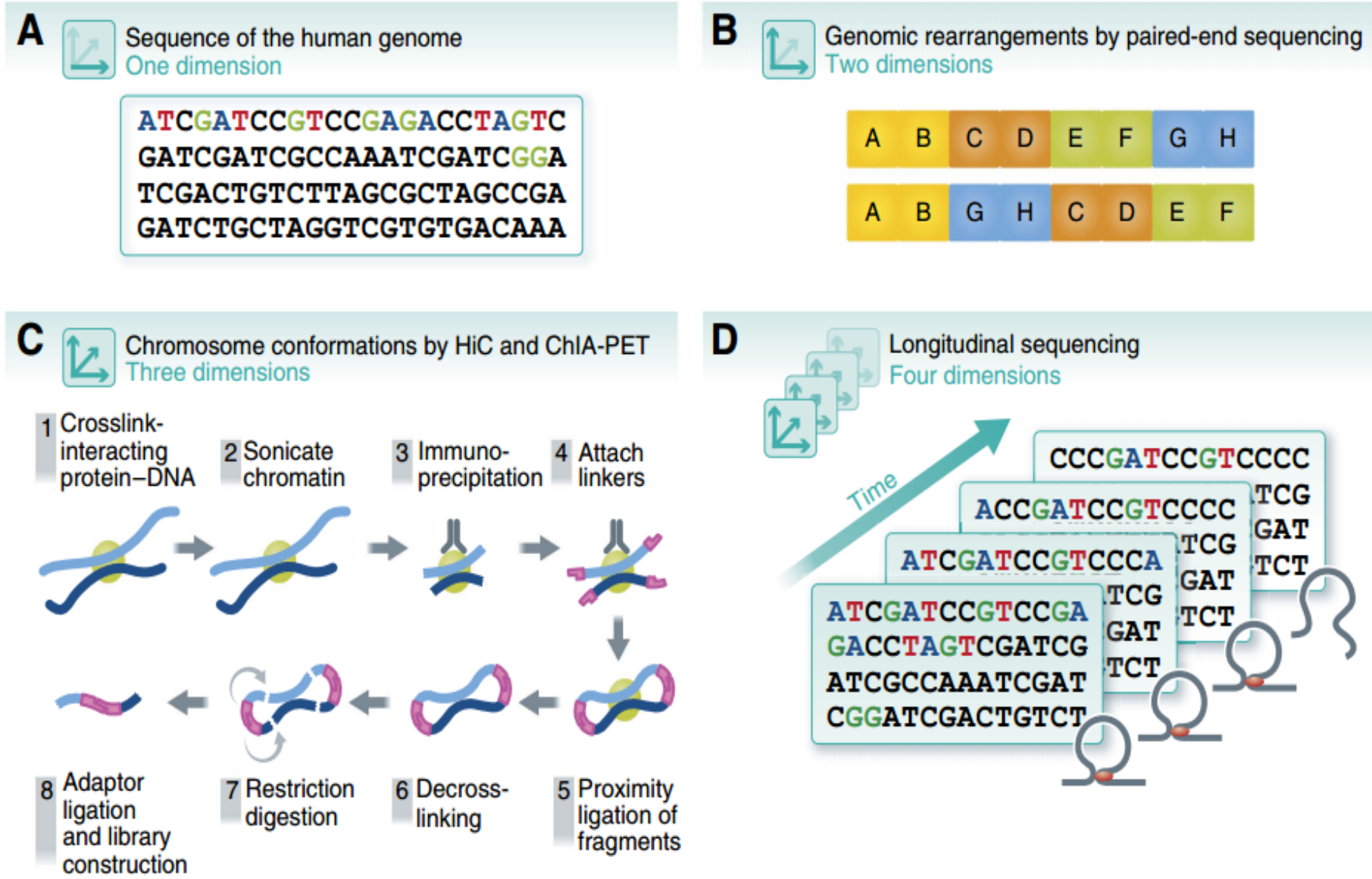

Figure 1 Dimensionality of the genome. The understanding of the human genome has expanded with advances of sequencing technologies, from (**A**) 1D sequencing of the human genome to (**B**) 2D mapping of SVs using methods such as paired-end sequencing, (**C**) 3D genome-wide chromosomal conformation capture using ChIA-PET and Hi-C, and (**D**) four dimensions across time.

- 除了破译基因组序列，NGS也应用在绘制表观遗传标记物图谱，比如DNA甲基化、组蛋白修饰。
- DNA中胞嘧啶残基的甲基化作用是最熟知的表观遗传标记，它通过诱导染色体凝聚来沉默基因。DNA甲基化可以在多细胞分隔中稳定遗传，因此使它可以调控生物过程，比如细胞分化、特异组织转录调控、细胞鉴别和基因组印记等。眼癌、直肠癌、白血病乳房和卵巢癌都和抑癌基因启动子的超甲基化有关，临床试验中也表明甲基转移酶抑制剂在治疗严重的骨髓白血病中是有效的。
- 通过几种NGS技术可以精确绘制这些泛基因组的甲基化模式图谱，包括MeDIP（methylated DNA immunoprecipitation，DNA甲基化免疫共沉淀技术）、MethylC-seq（胞嘧啶甲基化测序技术）、RRBS（reduced representation bisulfite sequencing，简化的表观亚硫酸氢盐测序技术）。

DNA modification

- 组蛋白修饰是一种直接影响基因调控的表观遗传现象，其异常修饰导致基因失调类疾病，在183例前列腺癌中检测出5种组蛋白标记发现不同类型的组蛋白修饰带来不同风险的反复肿瘤。区别于用组蛋白修饰蛋白的抑制物，更直接的修饰位点和目标修复可能会是更有用和重要的治疗策略。
- 组蛋白修饰位点可以通过ChIP-seq得到，通过这种方法在CD4+T细胞中发现了39个组蛋白修饰，最近的一项计划在46种细胞中发现了12种类型的修饰作用，显示组蛋白修饰作用具有细胞特异性。

Histone modification

Content

- Genomes, variation and epigenomics
- Transcriptomes and other functional elements in genomes
- Medical genomic sequencing
- Single-cell sequencing
- Future developments

- 除了基因组测序和相互作用分析，NGS也使RNA-seq技术可以绘制全局转录组图。高通量技术可以检测并定量分析转录产物、发现异常的异构体和其与基因变异（特异的等位基因变异）表达间的联系。揭露各种调控因子在控制基因表达中的作用也具有重大意义，比如转录因子和非编码RNA。

Transcriptomes and other functional elements in genomes

- 生物芯片技术第一次使在转录水平测量全基因组成为可能，但其只能研究已知的基因并且在交叉杂交和高噪音水平表达上有严重缺陷，而RNA-seq使在mRNA水平上测量更准确，这种技术利用高通量cDNA片段测序建立完整或间断的RNA文库，这可以清晰的绘制特异的基因组区域而没有或有很少背景噪音。RNA-seq可以精确的定量转录物和外显子，也可以用来分析转录物的异构体、鉴别异常的DNA和RNA，这包括用末端配对RNA-seq辨认异常的融合基因表达和发现新的非编码RNA（如lncRNA， long non-coding RNAs ）。还可以用来绘制在真核细胞转录组中广泛存在的等位基因的特异性表达（ASE, allele-specific expression ）和编辑位点图谱。

Transcript detection and quantification

- 应用高通量的技术使检测和定量分析转录物在医疗中显得更重要，在缺血性中风和Ⅱ型糖尿病中发现异构体的特异性表达是有害的；Ⅰ型转化生长因子-β（TGF-β）受体的ASE在直肠癌中具有遗传倾向；DAPK1基因ASE和慢性淋巴白血病有关。在口腔鳞状细胞癌肿瘤细胞和正常细胞中比对会导致基因表达改变的等位基因失调，结果表明在癌症相关功能中这些基因得到加强，证实等位基因失调是癌症病原的基本因素。用RNA-seq技术绘制转录组图谱发现在黑素瘤和老年吃痴呆症中有几种特异的转录物和基因融合现象，这也显示NGS在解释人类疾病原理中的重要性。

- 转录丰度只是分析基因产物表达的一种手段，最近可以用GRO-seq（Global Run-On Sequencing）或NET-seq（Native elongating transcript sequencing）技术测初期RNA及其序列。可以用转录组动力学分析的方法描述转录物等的合成与衰减，这种尝试发现了和启动子相邻的暂停和激活基因，在肺纤维细胞中发现的类似激活基因数是用生物芯片技术发现的2倍以上。
- 蛋白质表达不仅在转录水平还在翻译水平受调控，Ribo-seq技术就是一种定量研究mRNA与核糖体结合状况的方法，而NGS的加入为多水平的基因表达提供了更广阔的检测视野。

Profiling transcript production and ribosome-bound mRNAs

- 大部分基因调控被认为发生在转录水平，因此转录因子连接位点也和基因表达调控紧密联系。ChIP-seq不仅提供高分辨率的转录因子结合位点，还可以在全基因组水平上绘制组蛋白标记图谱，这种技术已经应用在分析DNA绑定区域和鉴别个体间的表达差异。一些互补的技术也用来鉴别潜在的调控位点（包括DNA酶敏感位点），例如FAIRE(formaldehyde-assisted isolation of regulatory elements，甲醛辅助的调控元件隔离)、Sono-seq（Sonication of Cross-linked Chromatin Sequencing，超声波降解染色体交联测序）。
- 除了和染色体作用的蛋白质，一些RNA（如lncRNA）也是重要的调控因子并且这些RNA和很多生物过程联系。CHIRP（chromatin isolation by RNA purification）就是一种有效的检测RNA和DNA之间相互作用关系的高通量测序方法。

Genome-wide identification of protein–DNA interactions

Content

- Genomes, variation and epigenomics
- Transcriptomes and other functional elements in genomes
- Medical genomic sequencing
- Single-cell sequencing
- Future developments

- 基因组测序给医疗带来了巨大的影响，原来由于费用和通量的限制基因在临床医学上的应用不可实行，现在大约$5000和花费几天的时间就可以完成普通的人类基因组测序，这使得基因诊断在医疗上大规模应用，如X-连锁智力障碍、先天性糖基化紊乱、先天性肌肉营养不良、罕见的基因紊乱携带者或胎儿产前染色体诊断。当然使用的时候需要小心翼翼，因为很多基因可能会表现出假阳性或假阴性。

Medical genomic sequencing

- 癌症是一种基因疾病，癌症细胞基因组的高通量测序是了解这种复杂的基因疾病的主要方法，荧光测序、RNA测序、末端匹配测序和全基因组测序使我们了解越来越多的使身体反复病变的原因，如变异、扩增、删除和易位。通过末端匹配测序发现在乳房癌细胞基因组中将近一半的结构重排带来转录融合（fusion transcripts），而且大约44%的转录融合可能被翻译了，在1/3的乳房癌样品中都检测出了一种异常的RPS6KB1–VMP1融合基因，用末端配对的方法给癌症和正常细胞测序发现，正常细胞含有更多的染色体倒位、删除和插入；而癌症细胞则含有大量重复、易位、复杂的重排、SNVs和SVs。
- 癌症细胞的测序工作可让我们揭示癌变激活的路径、提供我们可能对治疗有效的信息。认真选择合适的细胞进行测序可能会让我们发现癌症细胞癌变的发展和变异过程。

Genome sequencing in cancer

- 全基因组和荧光技术测序可以有效的诊断出罕见的基因疾病并由此给出最优的个体治疗方案。这种方法需要同时检测病人的家属，从和病人基因相关的个体基因组对比中找出变异。现在已经有很多例子。当然有时候并不是一定能找出致病的变异，而是仅仅给出了一个可能引起病变的列表。所以，这里的瓶颈就是如何才能解释清楚基因变异和他们对人类健康的影响。

Genome sequencing for clinical assessment of ‘mysterious’ diseases

- 全基因组测序和转录组分析让变异和基因表达的改变变得清晰。但是直到最近基因组测序的力量才被人所熟知。而且多种不同测序技术的整合使我们可能接收的信息量变得更大。一项研究将个体的基因组全测序，然后进行超过14个月的后续的转录组学、代谢组学、蛋白组学分析。整合的分析带来了更加完整的个体基因组构造理解和患病风险评估，还追踪了Ⅱ型糖尿病的发生。广泛的研究揭示了多种生物系统功能是如何作用及从健康到患病时他们是如何一起改变的。这可能就是将来检测和诊断疾病的方法。
- 现在这种方案在医学中的应用仍然有很大的障碍，比如所需的时间、花费和工作量来完成如此大的数据分析，还有就是在追踪疾病的过程中花费的病人的利益。但是没有人可以否认这种方案在疾病治疗中的前景。

Personal genome sequencing for detecting medically actionable risks

Content

- Genomes, variation and epigenomics
- Transcriptomes and other functional elements in genomes
- Medical genomic sequencing
- Single-cell sequencing
- Future developments

- 生物研究往往涉及组织、细胞种群和整个有机体的分析。但是，大部分的变异却发生在个体细胞水平，这就使理解个体的性质在分析整个大系统中变得相当重要。就像癌症细胞，他们是由无性繁殖扩张而成的复杂群体，如果将它们作为一个整体来分析就会掩盖很多重要的肿瘤特点。这时就需要一种方案能够从部分或单细胞水平来分析系统特征。

Single-cell sequencing

- 到目前为止大部分测序技术要求从超过10^5个细胞中提取DNA或RNA，但这就是很大的问题，因为比如在肿瘤中存在复杂的成分，除了复杂的肿瘤细胞外，其中还夹杂着正常细胞（血细胞和纤维细胞），这就可能掩盖很多特征，而使分析结果变成了一个“平均数”。
- 激光显微切割捕捉技术（Laser capture microdissection）是一种获取“纯净”的肿瘤细胞样本的尝试。aCGH（array-based Comparative Genomic Hybridization，比较基因组杂交）可以在一个有限的分辨率和范围内分析单个细胞。
- 现在也发展了一些高通量、高分辨率的单细胞测序技术，这些平台将提高我们对癌症发生、恶化、转移过程的理解程度。这种更深的理解将在诊断、治疗中起到重要的作用。对于个人而言，也将提供更准确、细致的治疗方案。

Single-cell sequencing in cancer

- 像先前所说的，现在的基因组测序和转录组分析需要大量的细胞提供DNA或RNA，这就使在研究具有全能性和多功能性的细胞基因组表达和变异时相当困难。一个多功能的细胞到底是怎样分化成各种细胞、单个细胞又是如何“知道”自己应该如何分化都是很让人感兴趣的领域。
- 随着技术的发展我们可以分析单细胞的基因表达，现在人类64细胞的胚囊组织已经确定是由识别有差别的特异性标记物产生的。单细胞RNA测序技术可以更具深度和全面的分析全基因组水平的转录组分析，让我们知道更多在细胞内发生的事。但比这更多的过程仍然不为人知，分辨率和深度是现在最大的障碍。

Single-cell sequencing in embryonic stem cell developmental biology

Content

- Genomes, variation and epigenomics
- Transcriptomes and other functional elements in genomes
- Medical genomic sequencing
- Single-cell sequencing
- Future developments

- 虽然NGS给学科发展带来了巨大的贡献，但也存在很大的挑战。比如数据的存储，在过去的5年超过一百万人类基因组被测序，这是一个非常大的数据量（人类全基因组有3G），还有这些数据的安全问题，如何存储及谁有权限获得这些数据都使我们不能忽视可能带来的“基因歧视”；基因组注释也是很大的挑战，不仅需要分析基因组功能还要试图理解个体的变异和人类疾病间的关系。这些工作都带来了相当巨大的花费，虽然测序花费已经降低了很多但结合分析而带来的巨额花费仍然给应用在医疗上带来了极大障碍。巨大的数据量和复杂的分析工作也对计算机的计算能力提出了更高的要求。

Future developments

THANK YOU!