DOI: 10.13376/j.cbls/2023122
文章编号：1004-0374(2023)09-1115-05

睿●观●察

泛基因组：观察生物多样性和统一性的新视窗

吴家睿[1,2]
(1 中国科学院分子细胞科学卓越创新中心，上海 200031；2 上海交通大学安泰经济与管理学院，上海 200030)

世纪之交实施的人类基因组计划打开了人类认识生命的新途径。最为明显的改变是，研究者逐渐跳出了传统生命科学注重个别基因或蛋白质的“碎片化”研究模式，从整体或全局的角度来认识生物体及其生理和病理活动。*Nature* 杂志在 2008 年的一篇社论中对此变化有过一个确切的描述：“似乎在一夜之间就从一个基因、一个蛋白质、一个分子、一次研究一个，转变为所有基因、所有蛋白质、所有分子、一次研究所有。一切都按组学的规模进行”[1]。更为重要的是，研究者能够在基因组序列比较之基础上去探讨生命世界里不同生物的关系，从个体差异到群体特征。例如，2006 年启动的大型国际研究计划 “Zoonomia Project” 就是针对包括人类在内的数百种哺乳动物进行基因组测序，然后通过序列比较分析探讨这些物种之间的遗传变异，进而理解哺乳动物的演化规律，以及人何以为人。*Science* 杂志在 2023 年 4 月 28 日发表了 Zoonomia Project 首批 11 篇研究论文的专辑，包括了 240 种哺乳动物的全基因组测序和分析；该专辑的编者认为：“Zoonomia 计划预示着一个新时代的到来，即对数百个物种的基因组进行联合测序和比较研究将打开一扇大门，进而可以从全新的思路来理解哺乳动物、哺乳动物进化以及人类自身”[2]。可以说，比较基因组研究是 21 世纪生命科学领域一种利用基因组信息认识生命及其生理和病理活动的重要研究手段，其中最值得关注的就是“泛基因组”(Pan-Genome) 概念的提出和应用。

1 泛基因组是揭示生物多样性的“底层逻辑”

在 2001 年人类基因组草图发布之际，研究者给出的只是一个“参考基因组序列”(reference genome sequence)，大约 70% 的基因组序列来自同一个志愿者[3]。该参考基因组序列可以被用来代表人类作为物种的基因组特征；但是，来源如此单一的基因组参考序列显然不能够反映人类内部种族或个体之间的遗传多样性。

2005 年，美国研究人员在一篇链球菌基因组分析的论文中首次提出了“泛基因组”的概念[4]，并在同年发表的一篇题为“微生物泛基因组”的综述文章中讨论了这个新概念，即“一个细菌物种可以通过它的泛基因组来刻画；泛基因组由‘核心基因组’(core genome) 和‘非必需基因组’(dispensable genome) 组成，前者是指该物种的所有菌株中都存在的基因，而后者则指只在 1 个或 1 个以上的菌株中存在的基因”[5]。

由此可见，参考基因组只是代表“核心基因组”，而泛基因组还必须拥有代表种内群体遗传多样性的“非必需基因组”—— 相当于群体内个体基因组序列的总和。2009 年，中国研究人员在人类参考基因组序列的基础上整合了一个亚洲人和一个非洲人的基因组序列，首次实现了“人类泛基因组”的雏形[6]。目前，泛基因组概念和研究技术在微生物和动植物研究领域已经得到了广泛的应用，仅仅 2020 年就发表了 1 万多篇涉及到泛基因组的研究文章。

1.1 通过泛基因组认识生物体的基因组结构多样性

为了更好地研究人类的遗传多样性，美国和澳大利亚等多个国家的研究人员组建了“国际人类泛基因组参考联盟”(Human Pangenome Reference Consortium, HPRC)，其目标是“创建一个更复杂和更完整的人类参考基因组，一种基于图型 (graph-based) 的，从端粒到端粒组装的序列图谱作为全球基因组多样性代表”[3]。2023 年 5 月，HPRC 在

基金项目：中国科学院先导专项“多维大数据驱动的中国人群精准健康研究”(XDB38020000)；上海市科委“科技创新行动计划”软科学研究项目 (22692114600)

Nature 发布了首个人类泛基因组参考草图，包含了来自非洲、美洲、欧洲和亚洲多个国家不同祖源的47个个体的二倍体基因组序列；这个人类泛基因组参考草图比目前使用的人类参考基因组序列(GRCh38)增加了1.19亿个新碱基对和1 115个基因重复片段(gene duplication)；在新增的碱基对中有大约9 000万个属于结构变异(structural variant, SV)，如碱基片段的倒置、插入、缺失等[7]。需要指出的是，由于SV的复杂性，过去依靠参考基因组序列只能识别人类基因组中少量的SV；而现在基于人类泛基因组参考序列，SV的检测率可以提高104%[7]。此外，人类泛基因组参考序列的应用不仅能够提高SV的检测率，而且也能够提高对单个或数个碱基差异等基因组微小变异的检测率。在*Nature*同期刊发的另一篇研究论文中，美国华盛顿大学的研究人员利用人类泛基因组参考序列和新算法发现了人类基因组中数百万个新的单核苷酸变异(single nucleotide variant, SNV)，且这些新发现的SNV主要位于基因组的“片段重复”(segmental duplication, SD)区域内[8]。

然而，在HPRC分析的47个不同祖源的基因组样本中，东亚人群的基因组样本只有4个，其中3个为汉族。这使得该人类泛基因组参考草图不能全面地反映以中国人口大国为代表的东亚人群基因组结构多样性。这个缺憾被我国科学家组建的“中国人群泛基因组联盟”(The Chinese Pangenome Consortium, CPC)的工作及时填补。2023年6月，CPC在*Nature*发布了中国人群泛基因组参考图谱，在涉及到的58个核心基因组样本中，包括了汉族和36个少数民族；CPC发布的图谱比人类参考基因组序列(GRCh38)增加了近1. 89亿个新碱基对和1 367个编码蛋白的基因重复片段[9]。需要强调的是，CPC从这些中国人群基因组序列中共鉴定出了近1 590万个微小变异(small variant)和78 072个SV，其中590万个微小变异和34 223个SV在人类泛基因组参考草图中未被报道过[9]。

从物种的角度来看，现实世界中有许多物种，尤其是人工培育的、由多个品系组成的物种，其遗传多样性的程度要比人类的高很多。而泛基因组研究策略显然是研究遗传多样性程度高的物种之得力工具。因此，泛基因组研究策略近年来重点拓展到了农业种质资源的遗传多样性研究。牛是畜牧业最重要的经济物种，仅人工饲养的牛就超过600个品种。不久前，以美国科学家主导的国际团队组建了牛泛基因组联盟(Bovine Pangenome Consortium, BPC)，致力于对这些人工饲养的品种和各种野生牛品种的基因组进行测序和组装，从而发展出更为完整的表征牛基因组多样性的泛基因组参考图谱[10]。家蚕是我国农业最具特色的经济物种，我国研究团队不久前通过对1 078份蚕品种的基因组测序，获得了目前国际上最大规模、最完整的家蚕泛基因组图谱，鉴定出4 300余万个单核苷酸多态性(single-nucleotide polymorphism, SNP)、930余万个插入/缺失(insertion and deletion, InDel)、340余万个非冗余SV和7 308个新基因[11]。

泛基因组研究在农作物以及植物领域也很普遍。更值得关注的是，多个植物泛基因组研究超越了物种层面，让研究者能够在属的层面上构建“超级泛基因组”(super-pangenome)。2022年7月，中国农业科学院研究者牵头的研究团队发布了稻属超级泛基因组图谱，包括了亚洲栽培稻核心品系、非洲栽培稻、普通野生稻和短舌野生稻；鉴定到了近16万个SV[12]。同年8月，华中农业大学牵头的研究团队发布了覆盖了玉蜀黍属全部物种的超级泛基因组图谱，为单个玉米基因组的3倍，其中约37%的序列是玉米基因组所没有的[13]。2023年6月，我国研究团队发布了包括野生、地方种和现代栽培谷子品种在内的狗尾草属(*Setaria*)泛基因组，由7万多个基因家族组成，并含有6 000万个SNP、670万个InDel和20余万个非冗余SV[14]。

1.2 利用泛基因组寻找生物体的基因功能差异

基因组结构的遗传多样性意味着不同种群或人群以及个体之间存在着潜在的生理和病理差异。利用泛基因组技术找到的遗传多样性越丰富，研究者对生命的功能差异的认识就越完整，进而更好地推动对人类遗传学和复杂疾病的研究。例如，研究者从人类泛基因组参考草图中发现了1 115个蛋白质编码基因的拷贝数变异(copy number variant, CNV)；与人类参考基因组序列“GRCh38”相比，这些CNV增加了0.6~4.4 Mb的基因序列；且这些CNV基因与人类健康高度相关，包括淀粉酶(amylase；4~10个拷贝)和β-防御蛋白(β-defensin；3~7个拷贝)[7]。又比如，中国人群泛基因组参考图谱表明，中国人群特有的SV中近50%的重叠分布在6 426个蛋白质编码基因区的上游和下游各100 kb，其中4 344个基因被长度超过1 kb的SV所破坏，在这些基因中最常见的功能富集与免疫功能相关[9]。

不同于线性的参考基因组序列，人类泛基因组是基于图形组装的多维度图谱[3]，有利于寻找含有SD等复杂结构的基因组区域内的微小变异。研究者通过人类泛基因组草图，从基因组SD区域内发现了数百万SNV，进而构建了IGC (interlocus gene conversion)的全基因组图谱，包括498个受体和454个供体热点，共影响大约800个蛋白质编码基因的外显子；这些基因中有38个编码基因被认为是“受限基因”，即这类基因在进化上是保守的，其突变往往会对生物体的生存或适应性产生有害影响；其中，一些受限基因如凝血因子Ⅷ或补体C4B与疾病有关[8]。

泛基因组研究策略还能够更好地促进遗传变异和表型变异的相关性研究。目前这方面常用的研究策略是基于SNP的全基因组关联研究(genome-wide association study, GWAS)，即将个体基因组序列与参考基因组进行比对，然后将得到的SNP与表型进行关联分析。但是，如果GWAS分析涉及到的表型没有在参考基因组上找到相应的功能基因时，就会出现GWAS定位区间与实际功能基因之间偏差较大甚至检测不到的情况。也就是说，传统的基因组分析技术在遗传变异和表型变异的关系研究方面存在着明显的空白，被称为“遗传度缺失”(missing heritability)[15]。如果采用拥有大量SV的泛基因组为参考基因组进行GWAS分析，就可以很好地解决这类因单一参考基因组而导致的遗传度缺失问题。在狗尾草属泛基因组的研究论文中，作者明确指出，在对某些性状的分析中，基于SV的GWAS显著提高了基于SNP的GWAS效率，且其中某些信号只能通过SV-GWAS检测到；此外，连锁不平衡分析显示，约36.9%的SV与相邻SNP (±50 kb)不相关，表明与SV相关的大量遗传信息未被SNP标记所捕获[14]。玉蜀黍属的超级泛基因组研究也得到相似的结论——相比于常用的SNP和InDel等遗传变异，SV能解释更多的表型变异，而且有37%的SV不能被已知的SNP或InDel标记所替代[13]。

由此可见，利用泛基因组研究能够获得单一参考基因组所没有的遗传变异，尤其是SV，进而能够识别出与这些遗传变异相关的表型变异，并可帮助基因定位和功能位点挖掘。例如，在番茄超级泛基因组的研究中，我国研究者对321个番茄群体中的SV和代谢物进行了GWAS分析，检测到17种果味挥发物和249种果实代谢产物的显著相关信号，表明SV与番茄的农艺性状变异和代谢物变化具有显著的相关性[16]。此外，研究者在玉蜀黍属超级泛基因组的研究中，发现一个SV与植株响应干旱胁迫相关；该SV是一个近2 kb的转座子插入，其插入位置位于目标基因上游的调控元件内，可能破坏了该元件与其转录因子的结合，导致目标基因在叶片细胞中的表达受到抑制，从而影响了植株的干旱胁迫响应[13]。

2　泛基因组是认识生物统一性的“直通车”

泛基因组正如一个钱币的两面，不仅能够反映出物种内的遗传多样性，而且能够表征整个物种的共性特征。这种“两面性”源自组成泛基因组的两个基本构件：可变的非必需基因组和稳定存在的核心基因组，前者是指只在部分种群/品系或个别个体中存在的基因，而后者则指在物种内所有群体或个体都存在的基因。因此，研究者构造泛基因组之目的不仅仅是用于研究遗传多样性，而且也可以通过泛基因组来认识群体的共性特征。正如HPRC所强调的：“本(泛基因组)计划的一个核心目的是，在泛基因组参考图谱中记录下人类基因组之间的遗传相似性和差异性”[3]。需要指出的是，在肿瘤研究领域，“泛”的概念正在从比较基因组层面推广成为一种研究肿瘤发生和发展的各种现象或规律的新策略。

2.1　通过泛基因组探寻群体的共性特征

人类遗传学研究表明，不同的种族/人群具有不同的遗传特征，揭示这些遗传特征将有助于对不同人种或人群的生理和病理研究。当前人类基因组学研究以及公共数据库的数据大部分来自欧美人群，非欧美人群的遗传学数据比较单薄；即使在最近发布的人类泛基因组草图中，亚洲人群的基因组信息的代表性也很不充分。中国人群泛基因组参考图谱的发布及时补上了这个空白。我国研究者通过对这两个泛基因组图谱的比较发现，中国人群基因组中存在特有的223个SV热点，涉及到807个蛋白质编码基因——这些基因显著富集在一些重要的功能通路上，如氧运输和血红蛋白结构。例如，珠蛋白基因簇的区域内存在两个中国人群特异性SV，包括一段20 kb的缺失序列和一段10 kb的重复序列[9]。显然，这一发现将为研究中国人群贫血症特有的致病机制提供全新的线索。此外，通过该泛基因组图谱鉴定到的一些中国人群特有的SV，显著富集在从东亚人群中发现的疾病相关变异中；这些疾病包括尿石症、肾结石和甲状腺肿，其相关

变异在一些亚洲地区非常普遍[9]。

基于泛基因组策略的共性特征研究同样被应用于动植物领域。例如，在家蚕泛基因组图谱的研究中，我国研究者鉴定到涉及家蚕育种的468个驯化相关基因和198个改良相关基因，并发现中国实用种和日本实用种之间只共享不到3%的改良作用位点，从而找到了这两种家蚕类型间能够产生强杂交优势的遗传基础[11]。实验室常用的小鼠近交品系大致分为两类，经典实验室品系和野生来源品系。2018年，欧美研究者对12个常用的实验室小鼠近交系和4个野生来源近交系进行了全基因组从头测序 (*de novo* assembly) 和分析，发现小鼠参考基因组中有2 567个区域表现出巨大的序列多样性，占小鼠基因组的0.5%~2.5%。这些多样性区域往往是品系特有的，被称作“品系单体型多样区”(regions of strain haplotype diversity)，主要是编码免疫、感知、神经、行为和有性繁殖等反映个体性状差异的相关基因。例如，野生来源近交系WSB/EiJ基因组在IRG、Nlrp1、Raet1等多个区域携带全新等位基因，与其他15个品系均不相同[17]。也就是说，尽管作为近交系的各小鼠品系之间的遗传背景高度一致，但比较基因组研究揭示出，不同品系之间依然携带着其品系特有的遗传特征。

对于在属内进行基因组比较研究的超级泛基因组研究而言，还可以比较同一个属内不同物种之间的特征和演化关系。例如，在稻属超级泛基因组研究中，研究者揭示了非洲栽培稻和亚洲栽培稻两个不同种之间对影响水稻株型的基因是如何进行独立变异的：野生稻种中存在一个由多个串联锌指基因构成的RPAD位点；在向栽培稻的演化过程中，为了适应其对应的环境，非洲栽培稻与亚洲栽培稻在RPAD位点两端分别发生了大片段缺失，从而各自拥有了RPAD位点两端不同的锌指基因，导致非洲栽培稻匍匐生长而亚洲栽培稻的株型变为直立[12]。在玉蜀黍属超级泛基因组的研究中，研究者注释了58 944个基因，并基于大的插入和删除变异——“存在 - 缺失变异”(presence-absence variants, PAVs) 鉴定出了“核心基因组”和“非必需基因组”，其中44.34%的基因是非必需基因；通过分析玉米野生近缘种“大刍草”(teosintes) 和现代玉米的691个基因型，重构了3个玉米亚群 (sub-populations) 和8个大刍草亚群，认为大刍草向玉米的驯化过程中可能同时发生了“老”基因的丢失和“新”基因的获取[13]。

2.2 采用“泛肿瘤”研究策略探寻肿瘤发生和发展规律

最能够体现出基于泛基因组的共性特征研究的工作是在肿瘤领域。肿瘤最主要的特点是，在基因组层面广泛存在着肿瘤患者个体间异质性 (inter-tumour heterogeneity)。为此，美国国立卫生研究院 (NIH) 在2006年牵头启动了一个国际合作项目——“癌症基因组图集”(The Cancer Genome Atlas, TCGA)，对33种肿瘤类型的上万名患者的肿瘤样本进行基因组测序和分析；在2015年该项目结束时，研究者已发现了近1 000万个肿瘤相关的突变[18]。需要强调的是，在TCGA实施过程中专门衍生出一个“泛癌图谱计划”(PanCancer Atlas project)，重点关注高度异质性肿瘤内部隐藏着的共性特征，要揭示控制癌症发展和进展的规律，正如2018年*Cell*杂志在刊发一系列“PanCancer”研究计划的论文时发表的社论所指出：“该图谱提供了一个独特而全面、深入和连贯的理解——肿瘤是如何、在哪里以及为什么在人类中出现的”[19]。

TCGA研究涉及到的33种肿瘤类型之划分是基于病理性状和解剖位置等传统肿瘤分类方法，但“PanCancer”研究计划则通过整合基因组、蛋白质组和其他组学数据，把这33种类型的肿瘤划分为28种整合分子群 (integrated clusters, iClusters)[20]。这种分子分型方法不仅重新界定了肿瘤的类型，而且有助于揭示不同肿瘤类型在分子层面的共同特征。此外，研究者采用“PanCancer”策略对TCGA中1万多样本的基因组数据进行致病性生殖细胞系突变 (pathogenic germline variants) 分析，发现了不同肿瘤类型之间的共同突变；具有代表性的是5种癌基因 (RET、AR、PTPN11、MET和CBL) 上携带的33种致病或可能致病的突变，其中21个RET基因突变存在于11种肿瘤类型之上[21]。

“PanCancer”研究思路还可以用于对肿瘤发生发展中的重要活动进行系统的比较研究。例如，研究者利用TCGA计划获得的基因组和转录组数据，对33种肿瘤类型近9千个样本基因组上的增强子 (enhancer) 进行分析，发现了这些肿瘤样本的一个共性特征：“基因组整体水平的增强子活性与非整倍体 (aneuploidy) 正相关，而与基因突变的程度则没有相关性”[22]。过去的研究发现，不同类型的肿瘤在进行远端转移时具有不同的特征，但对它们是否具有共同特征并不清楚。2019年，荷兰研究者采用“泛转移瘤”策略比较了20多种实体瘤的2 520

对转移性和原发性肿瘤样本的全基因组序列，发现这些不同类型的转移瘤中“全基因组扩增”(whole genome duplication, WGD) 程度都要比相应的原发性肿瘤高很多，前者的 WGD 平均值达到了 55.9%，因此，基因组内有程度高的 WGD 是这些不同类型的转移性肿瘤之共同特征[23]。

肿瘤患者不仅具有明显的个体间异质性，而且在个体同一肿瘤组织内还广泛存在着细胞间异质性 (intra-tumour heterogeneity, ITH)。显然，如何认识肿瘤细胞的 ITH 及其变化规律是研究者面对的一个巨大挑战。近年来发展起来的单细胞 RNA 测序技术 (scRNA-seq) 为研究细胞间异质性提供了新的“武器”。最近以色列研究者通过对大规模的单细胞 RNA 测序数据进行整合，形成了一个当前最全面的泛癌单细胞 RNA 数据库，从中揭示出 24 种肿瘤类型的 1 163 个肿瘤样本的转录 ITH 模式[24]。研究者进一步利用这些 ITH 数据去发现不同肿瘤中具有共性的 ITH 稳定表达程序 —— 论文作者定义为“元程序”(meta-program, MP)，鉴定出 41 种具有特定功能的 MP 类型；并将这种 MP 分析技术扩展到 6 种常见的非恶性细胞类型，通过这些 MP 绘制出肿瘤微环境中细胞与细胞之间的相互作用[24]。

这种“PanCancer”研究策略不仅用于分子层面和细胞层面的肿瘤生物学研究，而且被用于肿瘤临床数据的分析。美国研究者将 TCGA 计划涉及到的 33 种类型肿瘤的 11 160 名患者样本的临床数据进行整合，形成了一个“TCGA 泛癌临床数据源”(TCGA pan-cancer clinical data resource, TCGA-CDR)，包括了 4 种主要的临床结局终点 —— 总体生存期 (overall survival, OS)、无疾病期 (disease-free interval, DFI)、无疾病进展期 (progression-free interval, PFI)、疾病特定生存期 (disease-specific survival, DSS)；基于 TCGA-CDR，研究者为这些肿瘤类型中的每一种推荐了可以使用的临床结局终点[25]。由此可见，“PanCancer”的底层逻辑与泛基因组研究是一致的，即在确定生物体或生理病理活动差异性的同时从中提取出相似性，实现“个性”与“共性”之统一。

[参 考 文 献]

[1] To thwart disease, apply now. Nature, 2008, 453: 823

[2] Vignieri S. Zoonomia. Science, 2023, 380: 356

[3] Wang T, Antonacci-Fulton L, Howe K, et al. The Human Pangenome Project: a global resource to map genomic diversity. Nature, 2022, 604: 437-46

[4] Tettelin H, Masignani V, Cieslewicz MJ, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA, 2005, 102: 13950-5

[5] Medini D, Donati C, Tettelin H, et al. The microbial pan-genome. Curr Opin Genet Dev, 2005, 15: 589-94

[6] Li R, Li Y, Zheng H, et al. Building the sequence map of the human pan-genome. Nat Biotech, 2010, 28: 57-63

[7] Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. Nature, 2023, 617: 312-24

[8] Vollger MR, Dishuck PC, Harvey WT, et al. Increased mutation and gene conversion within human segmental duplications. Nature, 2023, 617: 325-34

[9] Gao Y, Yang X, Chen H, et al. A pangenome reference of 36 Chinese populations. Nature, 2023, 619: 112-21

[10] Smith TPL, Bichhart DM, Boichard D, et al. The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. Genome Biol, 2023, 24: 139

[11] Tong X, Han MJ, Lu K, et al. High-resolution silkworm pan-genome provides genetic insights into artificial selection and ecological adaptation. Nat Commun, 2022, 13: 5619

[12] Shang L, Li X, He H, et al. A super pan-genomic landscape of rice. Cell Res, 2022, 32: 878-96

[13] Gui S, Wei W, Jiang C, et al. A pan-*Zea* genome map for enhancing maize improvement. Genome Biol, 2022, 23: 178

[14] He Q, Tang S, Zhi H, et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. Nat Genet, 2023, 55: 1232-42

[15] Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature, 2009, 461: 747-53

[16] Li N, He Q, Wang J, et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nat Genet, 2023, 55: 852-60

[17] Lilue J, Doran AG, Fiddes IT, et al. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nat Genet, 2018, 50: 1574-83

[18] Ledford H. End of cancer atlas prompts rethink. Nature, 2015, 517: 128-9

[19] Kruger R. Charting a course to a cure. Cell, 2018, 173: 276-77

[20] Katherine A, Hoadley KA, Yau C, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell, 2018, 173: 291-304

[21] Huang KL, Mashl JR, Wu Y, et al. Pathogenic germline variants in 10,389 adult cancers. Cell, 2018, 173: 355-70

[22] Chen H, Li C, Peng X, et al. A pan-cancer analysis of enhancer expression in nearly 9 000 patient samples. Cell, 2018, 173: 386-99

[23] Priestley P, Baber J, Lolkema MP, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature, 2019, 575: 210-6

[24] Gavish A, Tyler M, Greenwald AC, et al. Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. Nature, 2023, 618: 598-606

[25] Liu J, Lichtenberg T, Hoadley K, et al. An integrated TCGA Pan-Cancer Clinical Data Resource to drive high-quality survival outcome analytics. Cell, 2018, 173: 400-16