目 录

1. 摘要

BGE（BGI Genomics Exploration Platform）是一个个人组学数据开放平台，致力于让组学数据走入每个人日常生活。BGE 平台基于先进的区块链技术，提供安全高效的个人组学数据分析、数据存储，SDK 及应用 API 接口。秉承“**测序一次，终生相伴**”的理念，赋能科研人员及合作伙伴，为用户提供丰富应用生态。

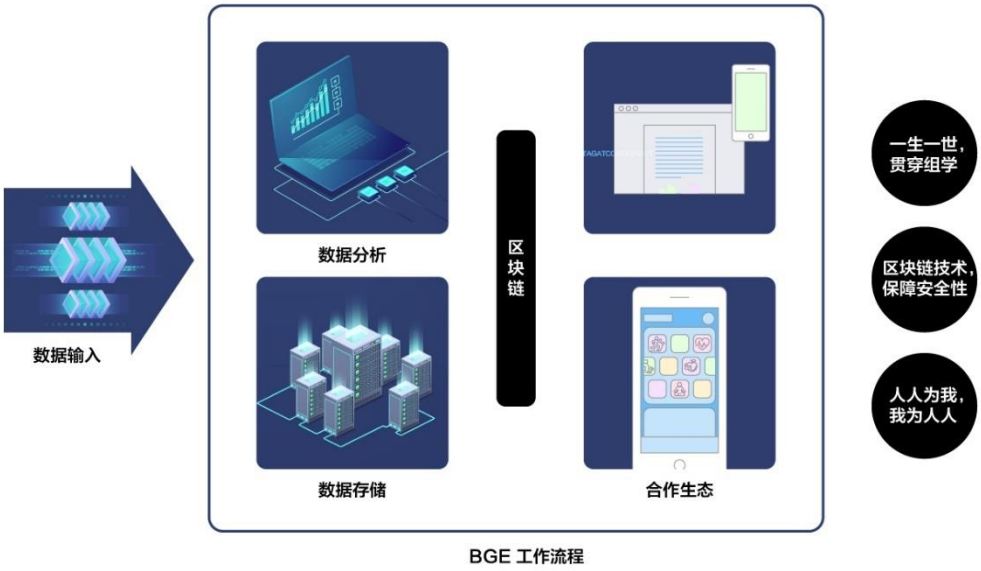

Figure 1 BGE 平台与生态

BGE 平台以个人组学数据的分析应用为核心，赋能科研人员及合作伙伴开发者，通过 BGE 提供的 API 和 SDK 来开发各种基于基因组及多组学数据的应用，例如：娱乐、科普、祖源、社交、风险评估、营养、体重管理等，从而实现用户“一次检测，终身相伴”的理念。

BGE 平台通过区块链技术，将数据控制权交给用户本人，实现个人数据细颗粒度授权与隐私保护，提升安全性。同时，平台支持用户以 300 万位点基因数

据及其他个人数据为基础，建立个人数据银行，实现个人数据资产化，帮助用户发现个人数据的巨大价值。

数据所有者通过区块链授权，许可个人组学数据进行科研研究。希望获得他人组学数据信息的个人开发者、研究机构或科研企业在得到所有者的同意后，可获得所需的组学数据资源。同时，应用开发者基于最新的科研进展和 BGE 提供的 API 和 SDK 来开发各种基于基因组及多组学数据的应用，让 BGE 用户（数据所有者）体验个人组学数据的巨大价值，实现 “我为人人，人人为我”。

2. 数据分析及标准

随着测序技术及多种生命数字化工具的突破，以大数据驱动的生命时代正在来临。生命是一个多层次的复杂系统，在时空中动态演变。华大基因创造性地提出了以“大人群生命组学大数据（2B4D）“的方法论来认知生命，即从 DNA 开始，遵循生命中心法则，从基因组到蛋白组到跨组学贯穿，从微观到宏观、从生到死的跨尺度、多维度、多模态、全方位、全周期的海量全景式生命大数据解读，这必将带来以基因组学为基础的个人生命跨组学大数据及其应用需求的爆发式增长。

目前，BGE 平台已经针对全基因组测序数据、宏基因组测序数据、表型问卷数据、临床体检数据等多组学数据搭建了数据仓库，并开发了 API，下面以全基因组数据和宏基因组数据为例，介绍 BGE 平台数据的整体情况。

2.1 全基因组数据

该部分文档通过展示 BGE 平台已产生的 1069 例全基因组数据的情况，为开发者提供数据完整性和数据质量的整体介绍。

2.1.1 检测方法介绍

检测实验室 用户委托深圳国家基因库数字化平台实验室选择 BGISeq 平台进行 Paired-end100bp 测序。深圳国家基因库数字化平台是一个集自动化、标准化、高产高效于一体的公共测序平台。

质量控制体系 深圳国家基因库数字化平台获得了 ISO 9001 质量管理体系，ISO 14001 环境管理体系、OHSAS 18001 职业健康与安全管理体系认证、ISO/IEC

27001 信息安全管理体系认证。

生物信息分析流程 全基因组测序数据下机后，参考团体标准 T/SZGIA 2—2018《人类全基因组遗传变异解读的高通量测序数据规范》[14]，根据其 GC 含量（40%~44%），总数据量（>=90G），Q30（>=80），平均错误率（<=2%）进行基础质控。对于满足质控的样本，使用 BWA-MEM（v0.7.17）进行比对，参考序列为最新版 GRCh38[12]。使用行业通用的 GATK（Genome Analysis Toolkit）（v3.8）HaplotypeCaller 依据官方推荐的最优流程进行变异检测，使用已知变异集 dbSNP Build 151，将覆盖深度小于 4X 变异标记为未检出，以确保平台给出的变异的准确度。

2.1.2 数据质量及验证

选取 BGE 平台已产出 1069 例全基因组数据进行数据完整性和重复性分析（男 495 女 574）。1069 例全基因组数据的平均测序深度中值为 43X，95%样本平均测序深度大于 35X。在比对参考序列后，95%以上样本在全部常染色体 98.63%的区域达到 4X 以上覆盖度，97.38%的区域达到 10X 以上覆盖度，71.16%的区域达到 30X 以上覆盖度。说明平台产生的数据在绝大多数样本上都能够提供全基因组上绝大多数区域的高质量的变异信息。

对于线粒体，在 95%样本上，平均覆盖深度在全部序列上为 4582X，说明平台数据不仅可分析线粒体常见变异，支持线粒体单倍型的分析和线粒体相关的单基因病检测，还可分析低频突变和线粒体拷贝数的检测；对于后者，芯片和全外显子数据还不能支持，体现了全基因组数据的优势。

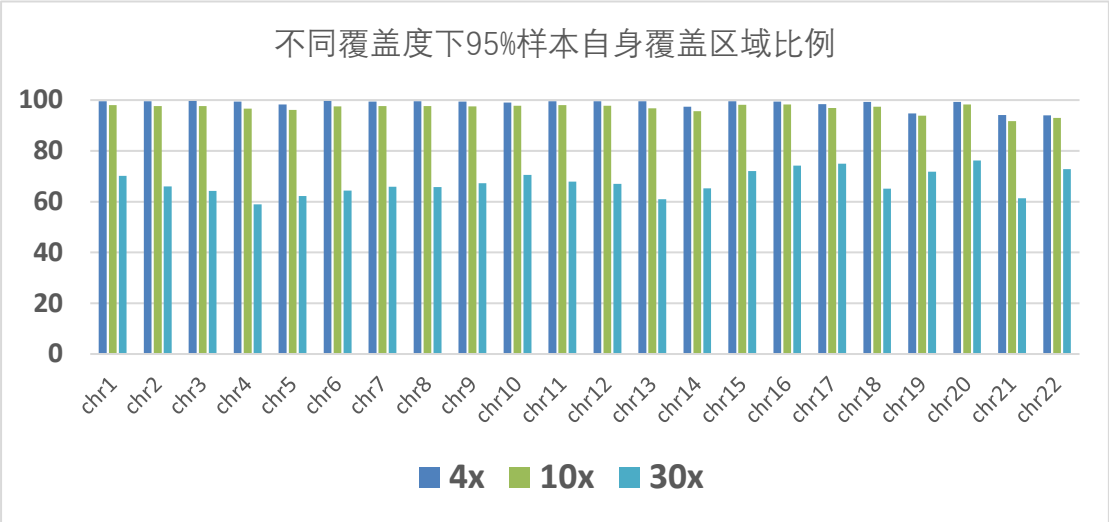

Figure 2 1069 例样本在测序深度为 4x、10x、30x 下 95%样本的统计

95%男性样本在 Y 染色体上达到 4X 覆盖度区域不低于 86.93%，达到 10X 覆盖度区域不低于 73.78%。由于对于男性，X 与 Y 染色体都是单倍体，预期覆盖度为常染色体的一半，因此可以根据平台提供的信息，检测 Y 染色体上的变异信息及进行 Y 染色体单倍型分析。95%女性样本在 X 染色体上达到 4X 覆盖度区域不低于 99.54%，达到 10X 覆盖度区域不低于 97.65%，95%男性样本在 X 染色体上达到 4X 覆盖度区域不低于 98.98%，达到 10X 覆盖度区域不低于 86.06%。

在所有染色体上，分染色体统计覆盖度超过 4X 和 10X 的区域所占的染色体中非 N 区的比例，结果如下表所示：

染色体	覆盖度% （深度>4x）	覆盖度% （深度>10x）
chr1	99.70	99.35
chr2	99.76	99.44
chr3	99.79	99.55
chr4	99.71	99.42
chr5	98.53	98.24
chr6	99.85	99.65
chr7	99.73	99.33

chr8	99.79	99.48
chr9	99.65	99.08
chr10	99.24	98.83
chr11	99.73	99.30
chr12	99.68	99.30
chr13	99.79	99.49
chr14	97.54	97.33
chr15	99.71	99.20
chr16	99.71	99.17
chr17	99.16	97.93
chr18	99.66	99.06
chr19	95.14	94.72
chr20	99.59	99.05
chr21	94.45	93.71
chr22	94.28	93.70
chrX（女性）	99.54	97.65
chrX（男性）	98.98	86.06
chrY（男性）	86.93	73.78

Table 1 染色体上 1069 样本不同覆盖度下 95%样本自身覆盖区域比例

准确度和灵敏度验证

为了评价全基因组数据的准确度和灵敏度，BGE 平台选择了国际公认标准品 NA12878[13]，使用与 BGE 平台相同的建库测序方案，进行了重复性实验。在平均覆盖度为 44X 的数据上，根据行业内公认的金标准数据参考集[2]，对 NA12878 检测到的变异准确度及灵敏度进行了评估。

对于单碱基突变（snv），在满足 4X 以上覆盖度的质控要求下，其准确度（Precision，平台检测出的单碱基突变和金标准给出的单碱基突变有多少比例相同）为 99.36%；灵敏度（Sensitivity，金标准单碱基突变有多少比率被平台检测

出）为 99.64%。对于小于 50bp 的短插入删除，在满足 4X 以上覆盖度的质控要求下，其准确度为 96.41%，灵敏度为 95.49%。

变异类型	总变异数	准确度	灵敏度
单碱基突变	30,533,92	99.36%	99.64%
短插入删除	243,912	96.41%	95.49%

Table 2 变异检测情况

2.1.3 数据应用场景

选取 2 个乳腺癌相关基因（*BRCA1,BRCA2*），2 个药物基因组（降脂药辛伐他汀）相关基因（*CYP3A4,SLCO1B1*）和 2 个较常见单基因疾病（苯丙酮尿和肝豆状核变性）基因（*PAH,ATP7B*）统计了其在 1069 个全基因组数据中的覆盖情况，可以看到全基因组数据在遗传性肿瘤基因分析、药物基因组分析、单基因病携带检测等应用领域均能达到良好的覆盖效果。

应用场景示例	基因	覆盖度% （深度>4x）	覆盖度% （深度>10x）
乳腺癌	*BRAC1*	99.89	98.77
	BRAC2	99.84	98.86
药物基因组	*CYP3A4*	99.97	99.73
	SLCO1B1	99.61	97.04
单基因病携带	*PAH*	99.72	98.49
	ATP7B	99.85	99.12

Table 3 1069 例数据在乳腺癌、药物基因组及单基因病应用场景的示例覆盖度统计

此外，从 GWAS Catalog（https://www.ebi.ac.uk/gwas/）中选取 2 型糖尿病风险相关位点（49 个），通过统计这些位点在 1069 个全基因组样本上的覆盖度，可以看出对于所有样本上全部 49 个位点，BGE 平台数据都有足够的测序深度来

支持得出准确的分型。相比芯片检测数据或全外显子数据，使用全基因组数据分析，可以让复杂疾病相关的科学研究及应用开发使用更丰富的位点，例如位于基因间区的位点；同时，在构建复杂疾病风险评估的模型中也无需考虑数据缺失的顾虑。

rs 编号	覆盖度%（深度>4x）	覆盖度%（深度>10x）	覆盖度%（深度>30x）
rs391300	100.00	100.00	100.00
rs11677370	100.00	100.00	100.00
rs9472138	100.00	100.00	100.00
rs5015480	100.00	100.00	99.91
rs10741243	100.00	100.00	99.91
rs10460009	100.00	100.00	99.91
rs4607103	100.00	100.00	99.91
rs4376068	100.00	100.00	99.91
rs4689388	100.00	100.00	99.91
rs896854	100.00	100.00	99.91
rs10906115	100.00	100.00	99.81
rs1387153	100.00	100.00	99.81
rs11634397	100.00	100.00	99.81
rs8042680	100.00	100.00	99.81
rs7630877	100.00	100.00	99.81
rs2383208	100.00	100.00	99.81
rs1552224	100.00	100.00	99.72
rs7578597	100.00	100.00	99.72
rs17036101	100.00	100.00	99.72
rs13292136	100.00	100.00	99.72
rs7901695	100.00	100.00	99.62
rs1531343	100.00	100.00	99.62
rs7957197	100.00	100.00	99.62
rs7172432	100.00	100.00	99.62

rs972283	100.00	100.00	99.53
rs17584499	100.00	100.00	99.53
rs5219	100.00	100.00	99.43
rs472265	100.00	100.00	99.43
rs243021	100.00	100.00	99.43
rs7578326	100.00	100.00	99.43
rs2943641	100.00	100.00	99.24
rs10923931	100.00	100.00	99.15
rs12027542	100.00	100.00	99.15
rs1359790	100.00	100.00	99.05
rs7178572	100.00	100.00	98.68
rs1048886	100.00	100.00	98.68
rs2237892	100.00	100.00	98.39
rs3773506	100.00	100.00	98.39
rs7961581	100.00	100.00	98.02
rs17045328	100.00	100.00	97.54
rs13266634	100.00	100.00	97.26
rs3792615	100.00	100.00	97.07
rs8050136	100.00	99.91	99.81
rs9460546	100.00	99.91	98.68
rs1153188	100.00	99.91	94.42
rs864745	100.00	99.81	98.49
rs6780569	100.00	99.81	93.10
rs12518099	100.00	99.72	93.48
rs5945326	100.00	99.53	85.54

Table 4 与 2 型糖尿病风险有关的 49 个位点在 1069 例样本上的覆盖情况

2.1.4 技术局限性

二代测序技术由于建库环节使用 PCR 扩增，对二级结构复杂的部分区域及 GC 含量偏高的区域和 polyA 结构，存在因扩增效率偏低而导致覆盖度相对较差的情况。针对二代测序技术这一局限，计划通过使用 PCR free 建库方法，对于测

序偏向性进行优化。

由于二代测序读长较短的局限，在全基因组范围内，对于高重复区域，例如短串联重复（str）无法给出准确的检测结果。对于传统的司法相关的应用，例如基因身份证，亲子鉴定等，无法直接使用原有的方法学，而需要使用基于单碱基突变的新方法。

对于遗传疾病的检测，例如亨廷顿氏舞蹈症，平台提供的分析结果仅作为科研数据参考，不具备临床级诊断效力，可通过结合 Sanger 测序的结果进一步验证。但对于基于 DNA 功能域(motif)长度估计的应用场景，例如端粒长度的估算，平台提供的数据目前不存在技术局限。

现阶段平台未提供结构变异及拷贝数变异的分析结果，针对这一局限，平台未来会通过结合单细胞 LFR（long fragment read）建库技术，使用独立开发的分析流程，给出更为准确的结构性变异和拷贝数变异的检测结果。同时对基因组进行定相（phasing），从而为遗传疾病的检测提供相比其他平台更为全面的信息。

2.2 宏基因组数据

人体共生微生物，指与人体相关的所有微生物集合，包括多种细菌、古细菌、真菌、病毒等构成的微生物群落。这些微生物群落，通过代谢，免疫和信号传导等多种途径，与人体的生理活动产生密不可分的联系，进而对人体健康产生影响。高通量测序技术的快速发展，帮助我们逐步揭开这些微生物群落的面纱，得以全面深入的了解人体这个超级生态系统。

考虑到现有研究基础和技术方案的成熟度，BGE 现仅为用户提供粪便样品来源的**肠道微生物检测**，包括样品采集、保存运输、实验室处理、高通量测序到

数据分析及存储等环节。人体其他部位的微生物样品在未来也会纳入服务范围。

BGE 采用宏基因组测序方法（Shotgun metagenomics sequencing）进行肠道微生物检测。宏基因组测序可全局非靶向地检测菌群中所有 DNA 序列，相比于 16S rRNA 等扩增子测序等方法，可更高分辨率地解析微生物群落的物种组成，并进行功能水平上的定量。BGE 采用业内通用的生物信息学分析方法，对菌群进行多维度的定量分析。以 API 形式，为开发者提供包括菌群中的基因、物种、功能的相对丰度信息，以及基于宏基因组数据产生的多种解读结果，同时，也提供各数据项对应的人群背景参考数据。

2.2.1 检测方法介绍

微生物组的宏基因组测序涉及样品采集和保存，DNA 提取、建库、测序等多个环节。BGE 使用业内领先技术，在不同的处理环节中减少 bias 的产生。以下介绍各环节中 BGE 采用的解决方案：

样品采集与保存 微生物组样品的采集与保存方法会很大程度影响到检测结果的准确性。BGE 采用 MGIEasy 粪便样本采集套装[1]进行粪便样品的采集，便于常温环境下的普通物流运输。

DNA 提取 BGE 选择 MagPure Fast Stool DNA KF Kit B (以下简称 MP)进行 DNA 提取。经"ZymoBIOMICS Microbial Community Standard" mock 标准品测试，MP 方法可很好保证细菌 DNA 完整度。DNA 质量控制标准按照深圳市团体标准,《基于高通量测序的环境微生物检测 第 2 部分：人粪便微生物宏基因组检测方法》(SZTT/SZGIA 1.2-2018)执行。

测序方法 BGE 选择 BGISeq 平台进行 Paired-end 100bp 测序。微生物组所

需的测序量，取决于菌群结构的复杂度和需要的分辨率水平。对于粪便样品，微生物组成结构相对复杂，BGE 使用高质量（Q20>90%, Q30>80%）的，去除人源数据的 $2.0*10^7$ paired-end reads, 4 GigaBases，用于下游分析。

检测实验室的质量控制 BGE 平台的用户样本送至深圳国家基因库数字化平台实验室进行微生物组样品的宏基因组检测。数字化平台获得了 ISO 9001 质量管理体系，ISO 14001 环境管理体系、OHSAS 18001 职业健康与安全管理体系认证、ISO/IEC 27001 信息安全管理体系认证。

生物信息分析流程 宏基因组生物信息分析流程包括，测序下机数据质控，去除宿主（人源）序列，基于参考序列比对的微生物组成定量（基因、物种、功能水平）。其中数据质控及去除人源污染使用 BGI 自有分析流程 [2]，基因水平的定量采用 IGC 肠道微生物基因集 [3] 为参考序列，物种定量使用 MetaPhlan2[4] 软件，功能水平定量使用 HUManN2[5] 软件。

2.2.2 数据质量及验证

稳定及可重复性 BGE 的样品处理流程，使用 6 个粪便样品，分别做了 6 次从 DNA 提取开始的技术重复验证，相关性分析显示，我们方法具有很高的稳定性以及可重复性。

定量水平	Spearman 相关系数	
	Average	SD
基因（gene）	0.8833	0.0471
物种（MetaPhlan2.species）	0.9005	0.0240
物种（mOTU）	0.9086	0.0371

Table 5 多次技术重复间的相关系数（Yang et al, 2018, unpublished）

检测结果的偏倚情况（bias）微生物组检测主要是对样品的组成情况进行定量分析，其中不可避免地在各个环节中引入误差，而其中 DNA 提取过程引入的偏差最大。我们使用"ZymoBIOMICS Microbial Community Standard" 标准品[15]评估了我们方法的偏倚情况，在对于其中细菌成分的定量，与理论值偏差在±20%以内，BGE 检测方法表现优秀(参照 SZTT/SZGIA 1.2-2018 标准[15])，对于其中的真菌成分，会稍微低估。由于粪便菌群主要是细菌，BGE 检测方法依然是适用于肠道菌群检测。

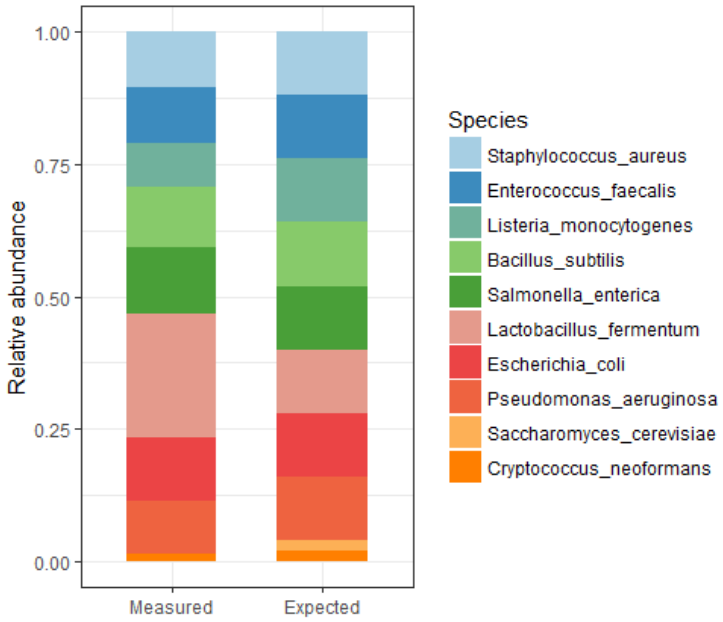

Figure 3 检测结果与 Zymo 标准品的比较 （Yang et al, 2018, unpublished）

测序数据完整度 我们使用完整肠道菌群基因集（integrated gene catalog, IGC[3]）对粪便宏基因组数据进行基因水平的定量，IGC 肠道菌群基因集包含约 10^7 个基因，可充分代表肠道菌群的基因。在 2,000 个中国人粪便的测试样品中，平均比对率达 80.7%，考虑到现有定量方法会导致部分数据在比对过程中丢失[3]，80.7%的比对率也接近理论上限（原核生物基因组中约 87%的区域为编码区）。这也说明宏基因组测序数据可完整测得肠道菌群中 DNA，IGC 参考基因集也适用

于中国人群。

数据库完整度 我们使用 MetaPhlan2 进行物种的定量，MetaPhlan2 的参考数据库包含~13,500 细菌和古细菌基因组，~3,500 个病毒基因组和 ~110 个真核微生物基因组的标志基因，已覆盖绝大部分人类已发现的微生物物种。

2.2.3 数据应用场景

已有大量的科学研究发现肠道菌群与人类健康的关系，也开始将肠道菌群数据，用于疾病诊断、个性化营养、个性化用药等健康管理的场景中，以下列出几个可能的应用场景：

健康风险评估 已经有不少肠道微生物的人群研究，使用统计学分析方法或机器学习模型（如，随机森林，SVM）[6,7]，发现了用来判别/预测疾病风险的 marker。开发者可基于 BGE 提供的数据，结合已有的数学模型，为用户提供健康风险预测服务。

个体化用药和个性化营养 由于肠道菌群结构的差异，对药物及营养物质的代谢有不同反应，比如，使用肠道菌群数据，预测人体餐后血糖的变化，从而选择合适的食物[8]；使用肠型分类，推荐使用不同的降糖药[9]。

连续监测 肠道菌群等人体共生微生物，是一个持续波动的、可干预的环境，尤其在婴幼儿时期，可通过连续监测肠道菌群的变化，评估婴幼儿的肠道健康发育状况[10]。BGE 鼓励开发者创建可连续监测用户健康状态的应用场景。

2.2.4 技术局限性

宏基因组测序技术应用于微生物组研究已有 16 年时间[11]，仍是一个高速发

展的领域，现有主要宏基因组数据应用的局限性包括：

生物信息学工具及数据库 现有宏基因组领域的生物信息分析工具及参考数据库还未有统一标准。虽然已经构建出高质量的肠道菌群参考基因集，由于缺乏高质量的注释参考数据库，其中仍有大量基因是未知功能和未知物种的，现基于肠道菌群的解读也主要基于有限的已注释信息。随着研究的进展和数据的积累，参考数据库会逐步丰富及提升可解释性。BGE 现选择最主流的几个菌群定量分析工具，为满足开发者需要，BGE 会保持数据库及相关生物信息工具的更新及迭代。

定量结果与真实环境的偏差 这个偏差由从采样到测序的多个环节引入，以至于不能完全正确反映样品情况。如上文所述，现有提取方法会低估样品中的真菌含量，虽然粪便样品中真菌含量相对较少，如若拓展至其他人体部位，BGE 也将选择更合适的提取方案。宏基因组测序是对环境中所有的 DNA 进行检查，也就包括了死亡裂解的微生物 DNA，所以检测解决不能反映取样环境中实际”活体”组成情况。此外，需要指出的是，这里检测的肠道微生物，是以粪便样品中的微生物为代表，而肠道的黏膜层仍有丰富微生物，受采样方式的局限，现没有便捷且无创的方法获得肠道黏膜层样品，这可能会对菌群与人体的免疫互作评估有较大影响。

人群肠道菌群异质性 我们也知道，人体共生微生物的组成存在巨大的人群和地理差异性。基于 A 人群建立的模型，不一定适用于 B 人群，这是在转化应用科学研究结果中需要注意的问题。

2.3 数据标准

为保证 BGE 平台数据的正确性、可用性和安全性，我们对数据资源进行标准化的管理，以解决生命数据语义存在大量的不兼容，数据格式多种多样，在数据收集、处理、存储和共享等方面缺乏统一标准等各种问题。

BGE 全系统适配华大金标准（Genomics Of Life Data Standard），该标准设计并制定了标准体系、数据字典及访问标准等，其中结构化数据的标准覆盖了个人信息、医疗、组学、环境等 8 个数据元集，目前已定义了共 6,626 个数据元（表 7 数据元属性），其中组学部分与 GA4GH 数据标准相互兼容，其他数据元集亦保持了对已有国标、团标最大程度的兼容性。

在数据进入 BGE 前，系统会对所有数据进行定义元及匹配元的操作，以保证数据在 BGE 或相关系统流动时有完全一致的数据元定义。为了方便开发者使用数据，我们还为 BGE 提供的数据内容根据其属性和涉及的知识领域设计了若干应用集，并将数据元集重新匹配为应用集，以提升开发者使用数据的便捷性及准确性。其中数据元为定义入仓数据的最小单位；数据元集指根据一般属性将相关数据元打包为集合；应用集指将数据元按照产品开发组织属性打包为集合。

序号	属性种类	数据元属性名称	约束	备注
1	标识类	数据元标识符	必选	专用属性
2		数据元名称	必选	专用属性
3		基础数据元标识符	可选	专用属性
4		基础数据元名称	可选	专用属性
5		合并数据元标识符	可选	专用属性
6		上一版本数据元标识符	可选	专用属性
7		外部数据元标识符	可选	专用属性
8		基础数据集名称	可选	专用属性
9		引用数据集名称	可选	专用属性

10		信息保护	可选	专用属性
11		版本	必选	公用属性
12		注册机构	必选	公用属性
13		相关环境	必选	公用属性
14	定义类	定义	必选	专用属性
15	关系类	分类模式	必选	公用属性
16		数据元值的数据类型	必选	专用属性
17	表示类	表示格式	必选	专用属性
17		数据元允许值	必选	专用属性
19		主管机构	必选	公用属性
20	管理类	注册状态	必选	公用属性
21		提交机构	必选	公用属性

Table 6 数据元属性

3. API 与 SDK

BGE 平台是一个跨越多级数据的（多）组学数据平台，开发者在获得用户授权后，将可以通过开放平台提供的 API 访问关联用户的（多）组学数据，甚至写入数据。基于开放平台 API 获取的（多）组学数据，使开发者避免采样、提取、测序和生信分析等繁琐漫长的过程。开发者可以通过开放平台提供的 API 创建极具创意性的（多）组学应用，如：

- 包含药物代谢基因的药物服用建议应用
- 包含体检数据和营养代谢基因的餐饮应用
- 包含肠道菌群数据和痛风相关基因的痛风管理应用
- 基于祖源成分的社交网络应用，等等

BGE 开放平台 API 使用 SSL 协议 + OAuth 2.0 标准，确保用户数据正确无误地授权用户需要的第三方应用，并安全无误地传输到第三方应用，确保用户数据不会被未经授权的个人或组织获得。

登录 https://open.bge.genomics.cn 可查阅更多内容。例如：关于 API 列表及调用方式，请浏览 “BGE 开发文档” 和 “示例开发 “。

4.用户隐私与数据安全

海量生命大数据的安全共享和高效应用是精准医疗的基础和生命时代的刚需，但数据的不合理使用通常会导致个人隐私泄漏风险。基因数据，特别是全基因组数据，具有高度敏感性，是具有全局唯一标识的生物特征数据，且蕴藏着大量多维度的敏感信息。例如人们可以从基因组数据中推断出关于种族、祖源、药物代谢能力、疾病风险、健康和行为、面部特征等信息，同时还涉及相关亲缘关系个体的遗传数据暴露，因此实现基因数据的完全脱敏匿名化异常困难。为了保护个人基因数据的隐私，全球各国政府积极制定了相关数据保护条例。例如美国的《遗传信息反歧视法案》(GINA)、《经济和临床健康信息技术法案》(HITECH)、欧盟的《通用数据保护条例》(GDPR) 等，均不同程度地限制了应用开发者的访问权限，最大限度地保护个人数据隐私。美国 NIH 和英国 Wellcome Trust 近来也更新了其数据共享政策，严格限制对单个基因型和聚合基因型频率数据的访问。

尽管隐私安全始终被执政者们放在重要位置，但近年来仍然发生了多起健康数据及个人隐私泄漏事件。2013 年，英国启动 Care.data 计划，旨在建立一个全国性的医疗健康大数据平台，以促进医疗和研究。但项目仅仅实施了三年，就被 NHS 叫停，原因在于该项目在共享和使用病人数据时，并未征得病人同意。

BGE 平台我们既需要切实地保护个人隐私，但同时也需要通过基因大数据的挖掘应用来推动研究进展、产品开发乃至整个行业与社会的进步。如何在组学数据应用的合理性与隐私保护间找到一个平衡点，保证数据共享对科研与社会的贡献超过其应用的风险，需要监管者、科学家、医疗人员和企业共同努力，建立基于伦理规范与技术解决方案的数据共享框架。

4.1 构建安全、高效的 BGE 生态社区

作为全球首个个人组学数据开放平台，BGE 将改变原有的基因组数据生产与分析模式，建立起数据所有者与应用开发者的桥梁链接，实现“测序一次，终生相伴”的目标。这既为个人基因数据的价值挖掘创造了实际应用场景，产生大量的数据申请、个人确权、应用反馈等用户交互行为，同时也将不可避免地导致对隐私安全与合规风险的担忧。

为了解决上述提到的数据安全问题，构建稳定、高效的 BGE 开放平台，我们以个人隐私保护为前提，以组学数据共享应用为目的，以价值激励生态为核心，前瞻性地融入了华大区块链 BaaS 平台为核心的数据流通生产级底层 IT 基础设施，为数据应用与安全保护提供了崭新的解决方案，其优势特点体现在以下几方面：

隐私保护：对标 GDPR 与国标 GBT 35273-2017 对个人数据的规定，为数据主体确权，实现个人数据的细颗粒度授权控制。具体包括：

(1) 支持对数据信息进行多重加密签名后的链上存储，确保平台上应用支付、数据调用等行为都如实、不可篡改地记录在区块链上，防止隐私泄漏；

(2) 支持凭证撤回功能，数据主体可监控数据应用，并对可疑侵权行为进行投诉与访问撤回，真正实现“我的数据我做主”。

安全共享：实现从数据仓库到 BGE 开发者之间的组学数据安全共享，确保数据共享全程可监管。具体包括：

(1) 支持安全多方计算平台，实现不同主体间相互不泄漏数据与算法前提下的数据协同计算，解决平台内用户之间数据共享的信任问题；

(2) 支持在云端进行密文分析，通过搭载可信计算环境，确保应用开发者

在无需下载或拷贝数据的前提下进行数据计算分析，切实保障数据安全。

价值交互：BGE 开放平台的愿景是通过对个人组学信息的挖掘，将数据资源资产化，形成生命价值可定价、可流通、可交换的全新生态体系。于数据所有者而言，建设基于通证经济体系的互动社区，实现个人确权下的数据价值激励反馈。于应用开发者而言，形成高效的数据应用模式与服务网络，极大降低了数据获取成本与违规风险，提高交易效率，最终形成一个多主体合作交互、多层次数据应用、多维度惠益分享的大数据应用生态体系。具体包括：

(1) 以个人所拥有的 300 万位点数据及其他个人数据为基础，以通证代币流通为模型，通过区块链技术建立个人数据银行与资产钱包，帮助用户发现个人数据的巨大价值并实现权益管理，量化个人对 BGE 生态的贡献价值；

(2) 通过遗传咨询、基因问答、家族社区等社交网络，在挖掘用户数据价值的同时，满足了用户对于知识解读、社群交友等功能的需求，提升用户粘性。

为了鼓励数据所有者与应用开发者积极参与 BGE 生态活动，通过贡献个人组学数据、算法模型等方式实现价值挖掘与回报，BGE 平台将会为普通用户与 DApp 应用开发者提供相应的激励措施。

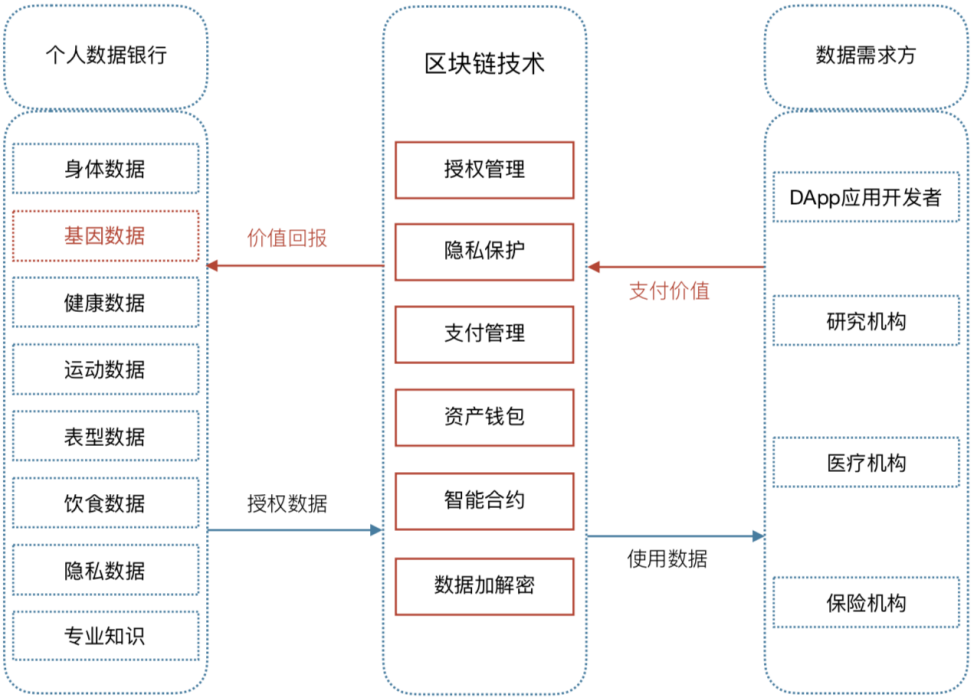

Figure 4 BGE 生态体系

BGE 生态应用场景示例

BGE 生态包含个人组学数据资产化应用、健康提升过程资产化应用、基因问答、遗传咨询、DApp 众筹开发与应用消费、游戏竞技等多维度应用场景。我们以个人组学数据资产化应用为例：

(1) 应用开发者生成算法模型并上传；

(2) 应用开发者发布测试邀请并向数据所有者提供相应的测试待遇；

(3) 数据所有者自愿参与本次活动，并授权个人组学数据给应用开发者进行计算；

(4) 数据所有者获得报告反馈，并结合个人表型特征，将结果准确性反馈给应用开发者；

(5) 应用开发者根据数据所有者的反馈对应用进行修改优化；

(6) 授权操作、支付操作等过程均通过区块链进行存证与结算。

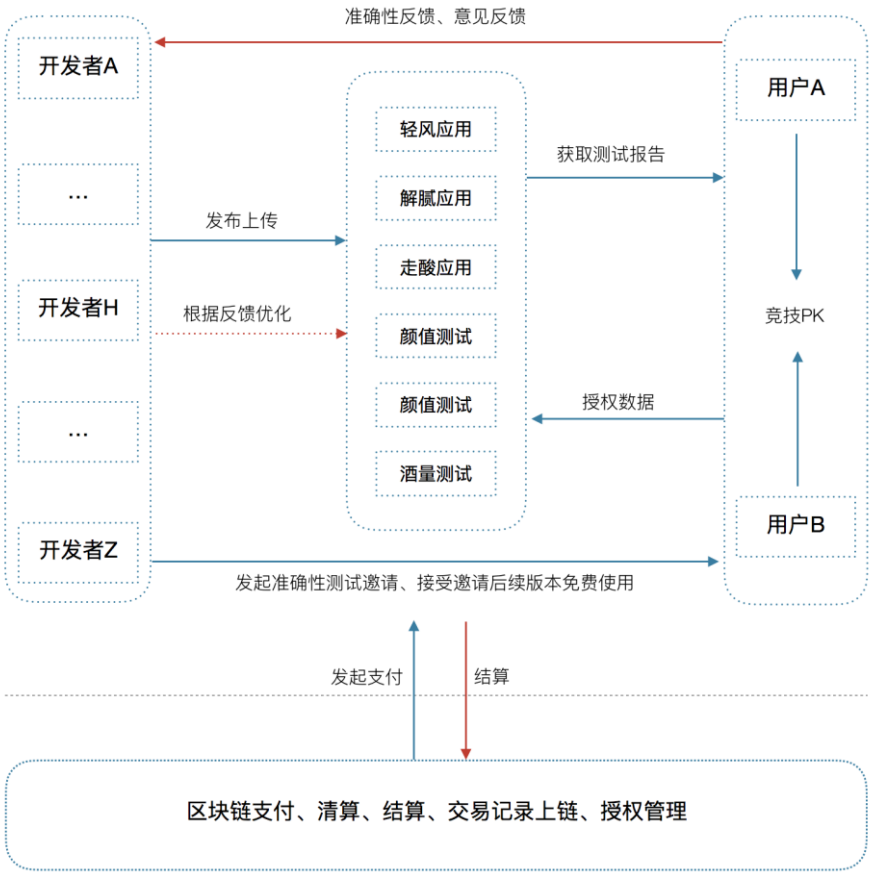

Figure 5 BGE 生态应用场景示例

4.2 基本数据安全

BGE 平台在每一层系统架构、每一个技术、甚至每一个系统操作或数据处理时都引入统一思考的信息安全管理体系。我们向第三方应用开发者提供基于 GA4GH 标准设计的 APIs 及符合相关隐私和安全策略要求的数据存储，保护，管理，访问，查询和传输的服务实体，确保数据在内部传输或上传到其他目的地时，符合适用于该数据和元数据的约束及语义规范。

为了保证数据的隐私和安全，我们对存储在同一物理设备上的不同 BGE 帐户的数据在逻辑上进行了隔离。只有一小部分 BGE 员工有权访问用户的数据。对于 BGE 员工而言，访问权限的分配是基于其工作职能，我们按照最小特权的

基本规则设计权限划分，并对她/他们的数据访问行为进行常规审计。对于需要额外访问数据的请求，将需要通过正式渠道向相应管理员申请并获得批准，在特定域及时间内进行访问，并将访问记录备份至信息安全部门。

我们在身份管理、授权管理、访问控制、隐私保护、审计日志、数据完整性、加密控制以及通信安全方面做了如下设计：

帐户授权/认证/访问控制

(1) 双因子认证：双因子认证为 BGE 帐户添加了额外的安全层，我们要求用户在输入用户名、密码时，同时输入验证码登录。这将降低 用户密码泄露时，未经授权访问发生的风险。验证码是通过短信运营商下发给用户预登记手机的一次性验证码。系统将在检测到用户登录域风险高时自动启用。

(2) 单点登录：单点登录（SSO）允许 BGE 帐户在多个应用中，只需要登录一次就可以访问所有相互信任的应用。我们集成了 CAS（Central Authentication Service）中央认证服务。

(3) OAuth 2.0：OAuth 2.0 是一个用于身份验证和授权的开放协议。BGE 的这项功能允许用户在不泄露用户名及密码等敏感信息的情况下登录到其他应用，并向其它应用授权其存放在 BGE 上的、范围内的数据访问。

隐私保护/审计/数据完整性

(1) 敏感信息存储加密：我们使用 Salted SHA-256 以及动态 AES-256 对敏感信息进行加密存储，以保证其在收集、使用、分享和再利用等阶段

符合法律和信息安全管理体系的要求。

(2) 敏感信息传输加密：BGE 在公开域进行数据传输时使用 HTTPS 加密（也称为 SSL 或 TLS）。BGE 服务器支持 ECDHE 密钥交换及 RSA 签名，这种 PFS 方法有助于保护传输中的数据，并最小化密钥泄露或密钥破解带来的安全风险。

(3) 关键系统日志审计：相关系统日志统一备份至信息安全部门，保存至少 1 年，并且 3 个月内的日志可实时分析，以便信息安全人员监控潜在的安全威胁和数据误用的活动，确保能将访问和披露活动追溯到单个帐户。

(4) 数据完整性 :BGE 会对特定域外的数据传输行为启用 SHA-2 验证，以确保其未在传输过程中遭受破坏。

5. 小结

BGE 个人组学数据平台，基于先进的区块链技术，提供安全高效的个人组学数据分析、数据存储，SDK 及应用 API 接口。秉承 “**测序一次，终生相伴**” 的理念，希望携手科研人员及合作伙伴，为用户提供丰富的应用生态。

参考文献

1. Han, M. *et al.* A novel affordable reagent for room temperature storage and transport of fecal samples for metagenomic analyses. *Microbiome* 6, 1–7 (2018).

2. Fang, C. *et al.* Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. *Gigascience* (2017). doi:10.1093/gigascience/gix133

3. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32, 834–841 (2014).

4. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903 (2015).

5. Huttenhower, C. HUMAnN: The HMP Unified Metabolic Analysis Network Harvard School of Public Health. 1–8 (2012).

6. Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nat Rev Microbiol* Accepted i, 508–522 (2016).

7. Pasolli, E., Truong, T., Malik, F., Waldron, L. & Segata, N. Machine learning meta-analysis of large metagenomic datasets : tools and biological insights. *PLoS Comput. Biol.* accepted, 1–26 (2016).

8. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* 163, 1079–1095 (2015).

9. Gu, Y. *et al.* Analyses of gut microbiota and plasma bile acids enable stratification of patients for antidiabetic treatment. *Nat. Commun.* 8, (2017).

10. Bäckhed, F. *et al.* Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 17, 690–703 (2015).

11. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35, 833–844 (2017).

12. Valerie A. Schneider1, Tina Graves-Lindsay,etc. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Research. doi/10.1101/gr.213611.116.

13. Mak, Sarah Siu Tze, et al. "Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing." GigaScience 6.8 (2017): 1-13

14. 团体标准 T/SZGIA 2—2018《人类全基因组遗传变异解读的高通量测序数据规范》, 深圳基

因产学研资联盟

15. 团体标准 SZTT/SZGIA 1.2-2018 《基于高通量测序的环境微生物检测 第 2 部分：人粪便微生物宏基因组检测方法》，深圳基因产学研资联盟