服务于精准医学的罕见疾病多组学信息注释平台

贾瑾萌¹, 明月¹, 安仲新¹, 梁云翔¹, 郭东铭¹, 李鑫¹, 陈庚¹, 田剑², 冯国双², 金雅琼², 郭永丽^{2*}, 倪鑫^{2*}, 石铁流^{1*}

1. 华东师范大学生命科学学院, 生命医学研究所, 上海市调控生物学重点实验室, 生物信息学中心, 上海 200241;

2. 国家儿童医学中心, 首都医科大学附属北京儿童医院, 北京 100045

* 联系人, E-mail: tlshi@bio.ecnu.edu.cn; nixin@bch.com.cn; guoyongli@bch.com.cn

收稿日期: 2018-06-15; 接受日期: 2018-07-03; 网络版发表日期: 2018-09-11

国家自然科学基金(批准号: 31671377)、北京市医院管理局重点医学专业发展计划(批准号: ZYLX201508)、北京市卫生系统高层次卫生技术人才学科骨干项目(批准号: 2015-3-079)资助

摘要 生物医学技术的发展, 以及近几年“精准医学”概念的提出, 使得人们对疾病的探究越来越多地深入到分子层面。与此同时, 医学诊疗也逐渐趋于个性化。作为发病率低、发病机制复杂、治疗难度极大的一类疾病, 罕见疾病越来越受到人们的关注。研究发现, 详细的临床表型是破译基因和实现罕见疾病的精准药物的基石。为了服务于精准医学的发展、辅助提高罕见疾病的诊疗水平, 我们建立了中文的罕见疾病多组学信息标准化平台eRAM (Encyclopedia of Rare disease Annotation for Precision Medicine, www.unimd.cn/eram)。该平台系统地整合了目前可获得的罕见疾病临床表现和分子机制的数据, 揭示了许多疾病之间的新关联。eRAM为15942种罕见疾病提供了丰富的注释, 包含6147种人类疾病相关表型、31661种哺乳动物表型、10202种来自UMLS的症状、18815种基因和92580种基因型。eRAM同时提供疾病注释体系、疾病网络、疾病预测、病例提交等功能, 不仅可以提供有关罕见疾病机制的信息, 还可以促进临床医生对罕见疾病做出准确的诊断和治疗决策。

关键词 罕见病, 疾病注释, 临床表型, 精准医疗

随着近年来生物医学技术的快速发展, 以及基因组测序技术、蛋白质组学、表型组学等各类组学技术的发展, 人们对疾病的病理病因在生物分子层面有了更加深入的了解。在此基础上, “精准医学”的计划成为疾病研究的一个重要发展方向。“精准医学”这一概念是由美国总统奥巴马率先提出的。2016年起, 我国也开始启动了针对各类疾病的“精准医学”项目。“精准医学”简单来说, 就是考虑个体变异性的疾病预防和治

疗策略^[1]。精准医学为疾病的治疗带来了新视角。然而, 精准医学的实施仍然存在一定的困难, 尤其是在开展个性化的医疗诊断上, 如何准确地将每个病人的临床表现信息(如临床表型、症状等)与分子水平的信息(如基因型、蛋白水平信息等)进行有效的对接成为开展精准医学的一个阻碍^[2,3]。

在实施“精准医学”计划的过程中, 最难系统开展的一类疾病就是罕见疾病。罕见疾病多为遗传性疾病,

引用格式: 贾瑾萌, 明月, 安仲新, 等. 服务于精准医学的罕见疾病多组学信息注释平台. 中国科学: 生命科学, 2018, 48: 1026–1032

Jia J M, Ming Y, An Z X, et al. A multi-omics annotation platform for rare disease to benefit precision medicine (in Chinese). Sci Sin Vitae, 2018, 48: 1026–1032, doi: [10.1360/N052018-00149](https://doi.org/10.1360/N052018-00149)

个体性差异大, 发病率极低. 由于缺乏可识别的症状, 临床医生往往无法对其做出最终诊断, 因此通常采用精准医学来根据患者的遗传内容选择最佳疗法. 虽然罕见疾病的患病率极低, 但是罕见疾病在世界范围内并不罕见^[4,5]. 我国作为一个人口大国, 罕见疾病群体数目更是十分庞大. 然而, 我国对于罕见疾病的研究无论是从信息管理与信息交流层面还是科研水平和医学诊疗层面都仍处于起步阶段. 因此, 建立一个标准化的罕见疾病体系是相当紧迫和必要的.

为了更好地服务于精准医学, 同时有效地辅助罕见疾病的诊断及治疗, 我们建立了标准化罕见疾病登记注释及交流平台 eRAM, 该平台可以为科研人员、医生、病人等关心罕见疾病群体的社会各界人士提供更加专业的帮助.

1 罕见疾病研究现状

根据世界卫生组织(WHO)的定义, 患病人数占总人口数的0.65%~1%的疾病统称为罕见疾病. 在我国, 对于罕见疾病的研究尚处于起步阶段, 目前并没有对罕见疾病做出明确定义. 2010年5月, 中华医学会遗传分会在我国罕见疾病定义专家研讨会上把罕见疾病定义为: 患病率低于五十万分之一, 新生儿发病率低于五万分之一的遗传病^[6]. 除了疾病定义不清的问题, 我国目前也尚未有官方的中国罕见疾病名录, 这就给罕见疾病的数据收集和记录造成了一定的困难. 此外, 与欧盟、美国等国家很早就开始建立罕见疾病的登记系统相比, 由于人口基数大, 医疗体系仍不够完善, 我国罕见疾病的研究还缺乏标准化的登记及注释体系, 而且我国至今还未建成标准化的罕见疾病体系^[7,8]. 就科研及医学诊疗而言, 我国对罕见疾病的研究还相当匮乏, 尤其是在临床水平与分子水平的结合方面还有很大的发展空间.

随着公众对罕见疾病的认识不断提高, 对罕见疾病的探索大多致力于相关的临床前和临床研究. 例如, 新一代测序(NGS)已被用于鉴定导致罕见疾病(包括一些新型表型)的基因, 同时需要大规模的表型注释信息^[9,10]. 人类表型本体(Human Phenotype Ontology, HPO)^[11]旨在实现对人类表型的大规模计算分析, 其中包含11600个用于描述个体表型异常的术语^[12], 然而, 基因与表型之间的关联仅限于少数罕见疾病. 最近,

研究人员又提出了罕见疾病治疗开发的工业化, 以降低治疗成本^[13], 这就更需要集中基于各种数据库的专业知识和资源. 在这些情况下, 基于表型系统的疾病标准化迫切需要整合临床表型和症状.

2 平台数据库构建

首先, 作为罕见疾病平台的构建基础, 我们建立了一个标准化的罕见疾病数据库. 截至目前, 通过对不同的疾病数据库和文献中罕见疾病相关信息的搜集和整理, eRAM中共有包括不同亚型在内的大约16000种罕见疾病. 此外, 为了保证数据的互操作性, 我们还对数据库中的各个字段进行了标准化处理.

2.1 罕见病定义与收集

目前, 罕见疾病没有统一、广泛被接受的定义, 罕见疾病在不同人群中的患病率有所不同^[14]. 为了促进更好的定义、数据共享和罕见疾病的诊断, 需要不同国家之间进行长期的合作^[15]. 为此, 我们整合了来自四个国际权威数据库: Orphanet^[16]、MalaCards(罕见疾病类别)^[17]、NIH遗传罕见疾病(NGRD)(<https://sen-sorisease.info.nih.gov/>)和国家罕见疾病组织(NORD)(<https://www.rarediseases.org/>)作为疾病名称资源(图1). 同时, 为了提供每种疾病的定义或简短描述, 我们从一体化医学语言系统(Unified Medical Language System, UMLS)中的MRDEF.RRF文件、Orphanet、在线人类孟德尔遗传(Online Mendelian Inheritance in Man, OMIM)、疾病本体(Disease Ontology, DO)、GARD (Genetic and Rare Diseases Information Center)以及NORD (National Organization for Rare Disorders)中提取并整理了6322种疾病定义. 截至目前, eRAM中共收录15942个独特疾病概念. 这些疾病概念包含了2018年我国最新发布的《国家罕见病目录》中的所有疾病, 并对其中87种疾病进行了细化分类(基于分子亚型对疾病进行了扩充). 例如, 对于《国家罕见病目录》中的“肌萎缩侧索硬化(Amyotrophic Lateral Sclerosis)”, eRAM不仅收录了该疾病, 还提供了其相对应的“AMYOTROPHIC LATERAL SCLEROSIS 1, ALS1”, “AMYOTROPHIC LATERAL SCLEROSIS 8, ALS8”, “AMYOTROPHIC LATERAL SCLEROSIS 4, JUVENILE; ALS4”等11种疾病亚型, 每个疾病亚型对应着不同的突变基因.

图 1 疾病来源统计

2.2 临床表型信息收集

在过去几年中, 我们通过多种模式分类方法对信息进行标准化提取, 从已发表的医学文献和临床资料中广泛收集了罕见疾病的表型和症状, 为大多数罕见疾病提供了丰富的临床和分子注释, 并最终形成了罕见疾病注释系统(统计结果见表1)。eRAM包括了从以下来源获得的症状和表型信息: (i) 人类表型本体(版本2017)。我们从HPO中为eRAM中的所有罕见疾病术语(包括它们的同义词)提取疾病-表型(D-P)。 (ii) DO症状, 使用“有症状”关系。 (iii) Orphanet。 (iv) UMLS (2017AA) 疾病表现文件——MRREL.RRF。最终共有16944种表型/症状被映射到1756种疾病。为了确保结果的高度准确性和完整性, 我们又基于模式识别的文本挖掘方法^[18], 利用HPO中的表型术语及UMLS中的症状术语作为词典, 对从MEDLINE中下载的8488796篇文献摘要和774514篇文献全文(2010~2015年)进行了文本挖掘, 最终得到549679条疾病-症状对和321792条疾病-表型对, 极大地扩充了现有疾病表型和临床症状的注释。在此过程中, 我们还保留了词对所在的语句, 以便医疗人员或科研工作者等专业人士进行审核或查验。考虑到在人类疾病研究中, 模式生物的广泛应用, 我们又进一步通过同源基因和语义识别, 将人类表型(Human Phenotype Ontology, HPO)^[11,12]与哺乳动物表型(Mammalian Phenotype Ontology, MPO)^[19]进行了映射。

表 1 疾病-症状/表型对数统计

数据类型	数据来源		
	数据库整合	文本挖掘	重合数据
疾病-临床症状	26187	15886	5281
疾病-临床表型	61230	53814	6960
疾病-临床表现 (症状与表型)	87417	69200	12241

2.3 疾病基因和基因型收集

新一代基因测序技术已经发现许多新的表型同基因之间的关系, 这些发现更加强调了对人类致病基因, 尤其是罕见疾病致病基因注释的重要性^[9,10]。因此, 我们也对罕见疾病的相关基因进行了收集整理。我们从包括Orphanet, OMIM, UniProtKB, ClinVar, DIS-EASES(文本挖掘数据)和DisGeNET(CTD数据)在内的多个现有数据库收集疾病基因关联, 同时建立疾病共现(包括共病与并发症)网络, 通过ClinVar中的数据推断疾病-基因关联。之后将它们整合后分为3类: 人工审核的基因、文本挖掘的基因以及预测得到的基因。目前, 数据库中共有316311个疾病相关基因注释信息, 涉及18815个不同的基因, 注释到5593个罕见疾病。另外, 对于我们收集的所有基因, 还添加了相应的基因座信息。eRAM包含来自以下资源的基因型信息: (i) 现有数据库: DisGeNET, GWASdb, LOVD和PharmGKB; (ii) 来自北京儿童医院的数据。eRAM总共包含92580个基因变体。在查询疾病后, 用户可以通过点击“基因”或“基因型”按钮查看这些数据。

2.4 数据标准化

由于我国现阶段还没有一个标准的罕见疾病官方名录, 为了使我们的平台的信息更加标准化, 提高信息的流通性和互操作性, 我们对数据库中的疾病名称、临床表现、表型以及基因都进行了标准化注释。疾病名称部分, 我们主要以UMLS(2017AA)为标准, 将疾病的同义词统一对应到唯一的UMLS编码。在此基础上, 进一步进行了同现存通用词表、本体以及数据库的映射, 包括医学主题词表(MeSH)、ICD10、医学分类命名法(SNIMED-CT)、DO, OMIM, GARD和Orphanet数据库。在对临床症状的标准化方面, 我们利用UMLS中的MRREL.RRF文件, 提取了其中50551条标准化的疾病-症状词对作为词表, 并利用文件中的句法关系进

行文本挖掘. 表型的标准化则是以HPO为标准, 对表型进行了同义词的扩充后进一步进行了文本挖掘. 基因方面仍依照NCBI的标准编码标准, 将每一个基因都统一到其Entrez ID.

3 平台功能

为了使平台更具专业化和全面性, 我们提供了5项主要功能: 疾病注释体系、疾病网络构建及可视化、罕见疾病预测功能、罕见疾病交流平台、病例提交功能.

3.1 疾病注释体系

基于我们的数据库, 在平台中共提供了罕见疾病8个字段的注释信息包括: 疾病定义、疾病同义词、疾病临床症状、基因、基因型、映射(Xref.)、人类表型及其相关的哺乳动物表型. 这些信息不仅能够为研究人员提供罕见疾病调查研究的相关机制, 还能为临床医生对罕见疾病的诊断和治疗提供帮助.

罕见疾病的注释体系是开放性的, 不仅对用户开放查询, 同时也提供一个Wiki注释系统. 在进行疾病查询时, 有两种方式供用户选择: 精确查询和模糊查询. 模糊查询就是用户可以通过输入已知的全部或部分疾病名称进行查询, 精确查询则是可以选择通过疾病在不同的数据库或标准化词表中的ID号进行查询. 在Wiki系统中, 用户可以为我们的平台中未注释的罕见疾病提交注释信息, 也可以对平台中已有的注释信息提出修改意见, 这些信息上传后, 我们将有专业人士进行审核, 审核通过的注释信息将录入我们的数据库中.

3.2 疾病网络构建及可视化

为了最大化地利用罕见疾病的相关信息, 将临床层面与分子层面的信息关联起来, 帮助科研人员研究罕见疾病的分子机制^[20,21], 探究疾病之间的关联性, 辅助医疗工作者通过已知疾病对未知疾病进行诊疗, 我们平台构建了3种疾病网络: 基于表型的疾病网络、基于基因的疾病网络以及基于共病的罕见疾病-常见疾病关系网络, 并对疾病网络进行了可视化.

基于表型的疾病网络是通过疾病共享的表型构建的, 基于基因的疾病网络是通过疾病共享的基因构建的. 对于用户输入的任一疾病, 系统将基于表型(或基

因)的相似性计算出与该疾病相似性最高的10个相关疾病, 用户可以选择查看其中任意一个与输入疾病的疾病网络, 网络图中将展示该两个疾病共享的表型(或基因)以及它们分别拥有的其他表型. 除此之外, 用户还可以根据自己对准确度的要求指定数据源(基于人工整理的表型、经过模式识别的表型或是文本挖掘的表型)构建疾病网络. 如图2, 为进一步方便用户查看疾病网络, 平台还提供了疾病-疾病网络模块, 用户能够利用该模块查看任意两个疾病的疾病网络, 该网络也是通过表型相似性构建的.

基于共病的罕见疾病-常见疾病关系网络主要是通过共病关系将罕见疾病同常见疾病联系起来. eRAM数据库中的共病信息主要通过两种方式收集. (i)从电子健康记录(EHR)中提取. 我们从多重参数智能监测重症监护(MIMIC II)数据库中提取共病信息, 从中收集了34261种独特的疾病共病对, 这些信息也保留在eRAM中的共病部分. (ii)来自MEDLINE的文本共病信息. 我们首先整合了eRAM, DO和OMIM的疾病概念, 并获得了171938个疾病术语(包括同义词). 然后, 采用上述描述的基于模式的方法从MEDLINE数据库中进行文献挖掘, 包括8488796篇摘要和774514篇全文文章, 最终挖掘出142422个不同的共病对和356845个句子. 文本挖掘结果由北京儿童医院专家人工审核, 所有文本挖掘的句子以及它们的PMID都已存入eRAM.

3.3 病例提交

为了推进罕见疾病的检测和诊疗, 我们的罕见疾病平台还建立了一个病例提交的模块. 这一模块主要是针对病人和医生设计构建的. 目前, 平台数据库中已收集了部分已经发表的病例样本, 且在持续收集, 这些病例样本是免费开放的. 另外, 希望为罕见疾病诊疗的发展贡献力量的病人和医生, 可以通过病例提交模块向我们平台提交罕见疾病病例, 提交完全依照自愿原则, 病人可以在系统中自行设置其信息共享的范围, 我们将对病人的个人信息严格保密. 对于向平台提交病例的人士, 我们将根据所提供的资料提供免费的诊断帮助, 对于有科研意义的病例, 我们会进一步对相关疾病进行免费的检测, 利用最先进的分子诊断手段及科研技术解析致病原因, 为疾病的治疗提供重要的信息.

图 2 基于表型相似性的疾病关联网络

4 平台优势

目前，eRAM共有15942个罕见疾病，对其中许多罕见疾病还注释了不同的亚型，是迄今为止国内外罕见疾病覆盖面最广的罕见疾病数据库。平台从临床水平和分子水平两个层面上，给罕见疾病提供了8个字段不同程度的注释。同时，与权威词表和其他本体或数据库的映射，也使得注释内容更加标准化。平台不仅构建了疾病网络，还提供了网络的可视化，使疾病关系更加形象，方便科研人员查看。平台功能多样化，是医学诊断有力的辅助工具。平台设计了疾病预测功

能、疾病注释信息展示、疾病网络可视化等多个功能板块，从不同层面服务于针对罕见疾病的精准医学的发展。医务工作者不仅可以通过疾病预测功能有效地辅助临床诊断，还可以通过平台提供的论坛及时交换诊疗意见，同时，患者及家属也可以方便地在平台上寻求治疗帮助。

5 结果与讨论

如今，罕见病在世界范围内引起了很多关注。美国国立卫生研究院推出了用于罕见疾病研究的未确诊疾

病计划, 加拿大也提出了加拿大FORGE(Finding of Rare Disease Genes)计划^[22]. 除此之外, 英国也已经开展了10万个基因组项目, 其中主要关注罕见遗传疾病, 目的是将基因组学诊断引入主流医疗系统, 为患者和研究人员带来益处^[23]. 这些项目的实施都离不开对患者临床表型和症状的标准化记录. 此外, 精准医学的关键在于收集和分析来自不同个体的疾病信息. 为了实现这一目标, 需要采用结构化和标准化的数据库来确保正确记录基于患者的数据. 因此, 我们建立了服务于精准医学的罕见疾病标准化平台eRAM(<http://www.unimd.cn/eram/>)用以辅助罕见疾病诊疗^[24]的发展. 作为罕见疾病研究和诊断的综合平台, eRAM目前为15942种罕见疾病提供了丰富的临床和分子注释, 包括综合的6147人类疾病相关表型术语、31661种哺乳

动物表型术语、由UMLS标准化的10202种症状以及18815种基因和92580种基因型信息. 以上信息的详细整合过程及标准化流程已在之前的文章中进行介绍^[25]. 此外, 本文从国内罕见疾病研究现状及面临的难题出发, 添加了罕见病网络展示、罕见病临床辅助体诊断功能以及面向中文的临床病例共享功能, 并提供了标准化罕见疾病与我国最新发布的罕见疾病目录所包含的罕见病之间的关系对比. 后续eRAM会进一步添加罕见病标准化中文注释信息, 推动中文临床信息与疾病多组学信息的整合分析. 目前eRAM几乎覆盖了所有罕见疾病, 可提供包含基因型及表型在内的多组学注释信息. 与此同时, eRAM将成为探索罕见疾病潜在机制的宝贵资源, 进一步促进新型治疗药物的开发.

参考文献

- Collins F S, Varmus H. A new initiative on precision medicine. *N Engl J Med*, 2015, 372: 793–795
- Arnedos M, Vicier C, Loi S, et al. Precision medicine for metastatic breast cancer—limitations and solutions. *Nat Rev Clin Oncol*, 2015, 12: 693–704
- Mirmazami R, Nicholson J, Darzi A. Preparing for precision medicine. *N Engl J Med*, 2012, 366: 489–491
- Dodge J A, Chigladze T, Donadieu J, et al. The importance of rare diseases: from the gene to society. *Archives Dis Childhood*, 2011, 96: 791–792
- Farwell Hagman K D, Shinde D N, Mroske C, et al. Candidate-gene criteria for clinical reporting: diagnostic exome sequencing identifies altered candidate genes among 8% of patients with undiagnosed diseases. *Genet Med*, 2017, 19: 224–235
- 魏珉, 张瑞丽, 赵志刚. 罕见疾病当前国际政策及现状介绍. *药品评价*, 2010, 7: 51
- Georgalas I, Tservakis I, Papaconstaninou D, et al. Pseudoxanthoma elasticum, ocular manifestations, complications and treatment. *Clinical Exp Optometry*, 2011, 94: 169–180
- Uitto J, Bercovitch L, Terry SF, et al. Pseudoxanthoma elasticum: Progress in diagnostics and research towards treatment summary of the 2010 PXE International Research Meeting. *Am J Med Genet A*, 2011, 155A: 3176–3176
- Boycott K M, Vanstone M R, Bulman D E, et al. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet*, 2013, 14: 681–691
- Freimer N, Sabatti C. The human phenome project. *Nat Genet*, 2003, 34: 15–21
- Robinson P N, Köhler S, Bauer S, et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Human Genet*, 2008, 83: 610–615
- Groza T, Köhler S, Moldenhauer D, et al. The Human Phenotype Ontology: semantic unification of common and rare disease. *Am J Human Genet*, 2015, 97: 111–124
- Ekins S. Industrializing rare disease therapy discovery and development. *Nat Biotechnol*, 2017, 35: 117–118
- Jia J, Shi T. Towards efficiency in rare disease research: what is distinctive and important? *Sci China Life Sci*, 2017, 60: 686–691
- Mascalonzi D, Knoppers B M, Ayme S, et al. Rare diseases and now rare data? *Nat Rev Genet*, 2013, 14: 372
- Pavan S, Rommel K, Mateo Marquina M E, et al. Clinical practice guidelines for rare diseases: the Orphanet Database. *PLoS ONE*, 2017, 12: e0170365
- Rappaport N, Twik M, Plaschkes I, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res*, 2017, 45: D877–D887
- Xu R, Li L, Wang Q. Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature.

- Bioinformatics, 2013, 29: 2186–2194
- 19 Smith C L, Eppig J T. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm Genome*, 2012, 23: 653–668
- 20 Schofield P N, Hancock J M. Integration of global resources for human genetic variation and disease. *Hum Mutat*, 2012, 33: 813–816
- 21 Robinson P N. Deep phenotyping for precision medicine. *Hum Mutat*, 2012, 33: 777–780
- 22 Beaulieu C L, Majewski J, Schwartzentruber J, et al. FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *Am J Human Genet*, 2014, 94: 809–817
- 23 McGrath J A. Rare inherited skin diseases and the Genomics England 100000 Genome Project. *Br J Dermatol*, 2016, 174: 257–258
- 24 Griggs R C, Batshaw M, Dunkle M, et al. Clinical research for rare disease: Opportunities, challenges, and solutions. *Mol Genet Metabolism*, 2009, 96: 20–26
- 25 Jia J, An Z, Ming Y, et al. eRAM: encyclopedia of rare disease annotations for precision medicine. *Nucleic Acids Res*, 2018, 46: D937–D943

A multi-omics annotation platform for rare disease to benefit precision medicine

JIA JinMeng¹, MING Yue¹, AN ZhongXin¹, LIANG YunXiang¹, GUO DongMing¹, LI Xin¹,
CHEN Geng¹, TIAN Jian², FENG GuoShuang², JIN YaQiong², GUO YongLi², NI Xin²
& SHI TieLiu¹

1 The Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China;

2 National Children's Health Center and Beijing Children Hospital, Capital Medical University, Beijing 100045, China

In recent years, with the development of biomedical technology as well as the concept of “precision medicine”, much effort has been devoted to relevant preclinical and clinical research, leading to the molecular study of diseases. Meanwhile, the diagnosis and treatment of diseases are becoming more and more personalized. With low incidence, complicated pathogenesis and large difficulty in treatment, rare diseases have aroused public awareness. Latest studies have found that detailed clinical phenotype is the keystone of deciphering genes and precise treatment for rare diseases. To develop precision medicine and to improve the diagnosis and treatment of rare diseases, we built a standardized Chinese platform named Encyclopedia of Rare Disease Annotation for Precision Medicine (eRAM, www.unimd.cn/eram), which systematically integrates the clinical manifestations and molecular annotations of rare diseases and provides new connections among rare diseases. At present, the platform contains a rich annotation for 15,942 rare diseases, including 6,147 human disease-related phenotypes, 31,661 mammalian phenotypes, 10,202 symptoms from UMLS, 18,815 genes, and 92,580 genotypes. eRAM provides enriched disease multi-level annotations, as well as connections among rare diseases. Moreover, additional functions are provided, including disease prediction and EMR (electric medical records) submission for users, providing information about rare disease mechanisms and assisting clinicians to make accurate diagnosis and treatment decisions for rare diseases.

rare disease, disease annotation, clinical phenotype, precision medicine

doi: [10.1360/N052018-00149](https://doi.org/10.1360/N052018-00149)