研究报告

Research Report

# **水稻中过敏原蛋白全基因组预测及生物学功能预测分析**

陈淼林1，许杰1\*

1. 上海交通大学 生命科学技术学院，上海200240）

\*通讯作者，E-mail: [xujie3000@sjtu.edu.cn](mailto:xujie3000@sjtu.edu.cn)

**摘要** 世界范围内过敏症状的流行趋势日益严峻，水稻作为单子叶模式植物和人类主要粮食作物，对于水稻中致敏性物质的研究具有重要意义。但目前基于水稻全基因组水平预测和分析过敏原的研究尚缺少相关报道。本研究利用生物信息学方法在全基因组层面上，在水稻基因组数据库中和过敏原数据库中，共筛选和获得了657个过敏原基因/蛋白（属于109个蛋白家族），并进一步对这些候选的水稻过敏原基因/蛋白进行了表达模式、家族分类和功能预测分析；以及在进化上分析了水稻过敏原基因/蛋白家族成员的同源蛋白在低等到高等植物中分布情况，结果表明致敏性可能随着基因家族中成员的功能分化而变化；进一步结合水稻花粉和种子转录组芯片数据，细化分析和获得了148个水稻花粉过敏原和116个水稻种子过敏原，并在基因家族和功能分类上进行了对比分析。本研究结果初步明确了水稻过敏原基因/蛋白家族的进化特点、组织表达特性和生物学功能，为进一步研究水稻以及其他单子叶植物中的过敏原奠定了基础。

**关键词** 水稻；预测；过敏原；转录组；功能分析

**Genome-wide Prediction and Functional Analysis of Allergens in Rice**

Chen Miaolin1，Xu Jie1\*

( 1.School of Life Sciences and Biotechnology，Shanghai Jiao Tong University，Shanghai 200240

\*Corresponding author，E-mail: xujie3000@sjtu.edu.cn

**Abstract** Allergies have long been a major pandemic health problem worldwide. As Rice is monocot model plant and one of the important food crops, the research of allergens in rice is of high significance. However, systematic studies of allergens in rice on whole-genome level are still lacking. In this study, genome-wide prediction and screening of allergens were performed. Based on allergen databases and genome information of rice, 657 putative and known allergens were obtained by using bioinformatics eventually. Then, analysis on expression pattern, protein family, and putative function of allergens were performed. Meanwhile, taxonomic distribution of allergen homologues from lower plants to higher plants was analyzed, which implying that allergenicity of allergens may change along with functional specification of protein families. Furthermore, 148 pollen allergens and 116 seed allergens were screened by transcriptomic data in rice and then their protein families and functional classification were compared. These research have initially elucidated the evolutionary traits, expression patterns and biological function of rice allergens will provide new information to promote future study of allergens in rice and other monocots.

**Key words** Rice; Allergen; Prediction; Transcriptome; Functional analysis

过敏性疾病发病率在世界范围内不断上升，被世界卫生组织列为21世纪重点防治的三大疾病之一(Pawankar et al., 2013)。使机体产生过敏反应的抗原分子称为过敏原，其本质大多为蛋白质(Larché et al., 2006)。其中，花粉和食物过敏原是导致过敏性疾病的两种最主要的元凶，开展花粉和食物中过敏原的研究，对于预防过敏性疾病的发生和提升人类健康水平和生活质量极其重要。目前的研究表明：花粉过敏原主要来自禾本科植物、野草和乔木三个物种，其中，禾本科花粉引起的花粉过敏症是最常见的花粉过敏症之一，特别是欧美地区(Yang et al., 2015)；而食物过敏原分布广泛，主要包括动物性过敏原（如牛奶及奶制品、蛋类等）和植物性过敏原（如大豆、小麦，水稻等）(Yang et al., 2004)。

水稻是最重要的粮食作物之一，约占世界粮食总产量的1/4，是1/2以上人口的主食，在世界粮食安全中占有十分重要的地位。但目前已经有研究表明：水稻也会引发人们的过敏症状，且主要集中于以稻米为主食的亚洲地区。研究表明，在日本超过10%以上的人群对水稻存在程度不同的过敏反应，在马来西亚这一比率甚至高达30% (Gendeh et al., 2000; Kosugi . et al., 1992)。此外，水稻也是科学研究中重要的单子叶模式生物(Goff and Briggs, 2002)。因此，对水稻中的过敏原物质的研究，不但对减少人们的过敏症状会有所帮助，也会对于禾本科乃至于其他植物物种中的过敏原研究有着重要的借鉴意义。

在水稻过敏原研究现状方面，目前日本研究处于领先地位，已经鉴定出几十种水稻过敏原蛋白。例如水稻花粉过敏原Ory s 1，Ory s 12和水稻种子Ory s 14和Ory s GBSS I等(Enrique et al., 2005; Hirano et al., 2013; Takumi et al., 2000; Xu et al., 1995)。花粉过敏原Ory s 1为扩张蛋白家族，Ory s 12为前纤维蛋白家族而种子过敏原Ory s 14为脂质转移蛋白家族，Ory s GBSS I为淀粉合成酶。但鉴于目前通过使用实验手段来检测、筛选和鉴定过敏原物质，具有周期较长、费用昂贵和难度大特点，且随着基因组学研究的发展，和过敏原信息的积累，利用生物信息学方法来分析和筛选过敏原物质越来越显现出其优势，已经获得很多成功的例子。目前，应用得最为广泛的方法是基于序列的预测方法、基于模体（motif）的预测方法及基于支持向量机（Support Vector Machine，SVM）的方法(FAO/WHO, 2003; Saha and Raghava, 2006; Stadler and Stadler, 2003)。其中基于支持向量机的预测方法更加关注蛋白的整体特性（通过特征向量），通常能够获得更好的结果，而依据模型的预测也意味着模型的好坏非常重要。用于构建支持向量机模型的集合以及特征向量的选择至关重要。Wang等将基于最大相关最小冗余的特征筛选方法结合支持向量机的方法，也取得了很好的效果(Wang et al., 2013)。有研究表明支持向量机方法预测过敏原的结果优于基于序列和模体的方法(Cui et al., 2007)。但目前在水稻中仍鲜有对过敏原进行系统预测和生物学功能分析的报道。

为了全面预测和分析水稻中的潜在过敏源蛋白及其特征和功能，本研究通过生物信息学手段，使用两种过敏原预测方法并结合已知过敏原数据库信息，在已经拥有大量研究数据的单子叶模式植物水稻中进行了全基因组过敏原预测分析；并同时对这些潜在过敏原的表达模式、蛋白家族分类和功能进行了系统的分析；此外，通过同源蛋白序列对比和进化分析，进一步了解到了水稻过敏原蛋白在不同植物物种中的分布特征；最后结合水稻花粉和水稻种子的转录组芯片数据，细分出了水稻花粉过敏原和水稻种子过敏原，并对这两种高危过敏原的表达模式、蛋白家族和功能进行了比较分析。上述研究有望对水稻过敏原的特征和功能进行系统的了解，同时通过单子叶模式植物水稻中的过敏原分析，能够帮助预测和分析其他物种中的过敏原，从而推进预防和治疗植物特别是禾本科植物引起的花粉过敏症和食物过敏症。

**1研究结果**

**1.1 全基因组预测和筛选水稻过敏原的基因/蛋白**

我们从水稻基因组注释计划数据库（http://rice.plantbiology.msu.edu/）收集了水稻中所有蛋白质序列共41363条蛋白质序列，通过PREAL和WHO/FAO rule 1两种过敏原分析方法，我们在水稻基因组中鉴定到645个过敏原蛋白（图1）。同时，我们增加和收集了过敏原数据库（Allergome, [http://www.allergome.org/）中的已知水稻过敏源蛋白46](http://www.allergome.org/）中的已知水稻过敏源蛋白共46)个，其中34个已知水稻过敏原在预测数据中。因此，最终在水稻中，我们得到共657个蛋白组成的潜在和已知过敏原集合（为方便描述，后文统称为水稻过敏原）。此外，我们分别从水稻成熟花粉和种子的转录组学分析中，选去出过敏源蛋白集合中的水稻花粉过敏原和水稻种子过敏原集合（图1），其中花粉过敏原95个，种子过敏原63个，在花粉和种子中均表达的过敏原53个。

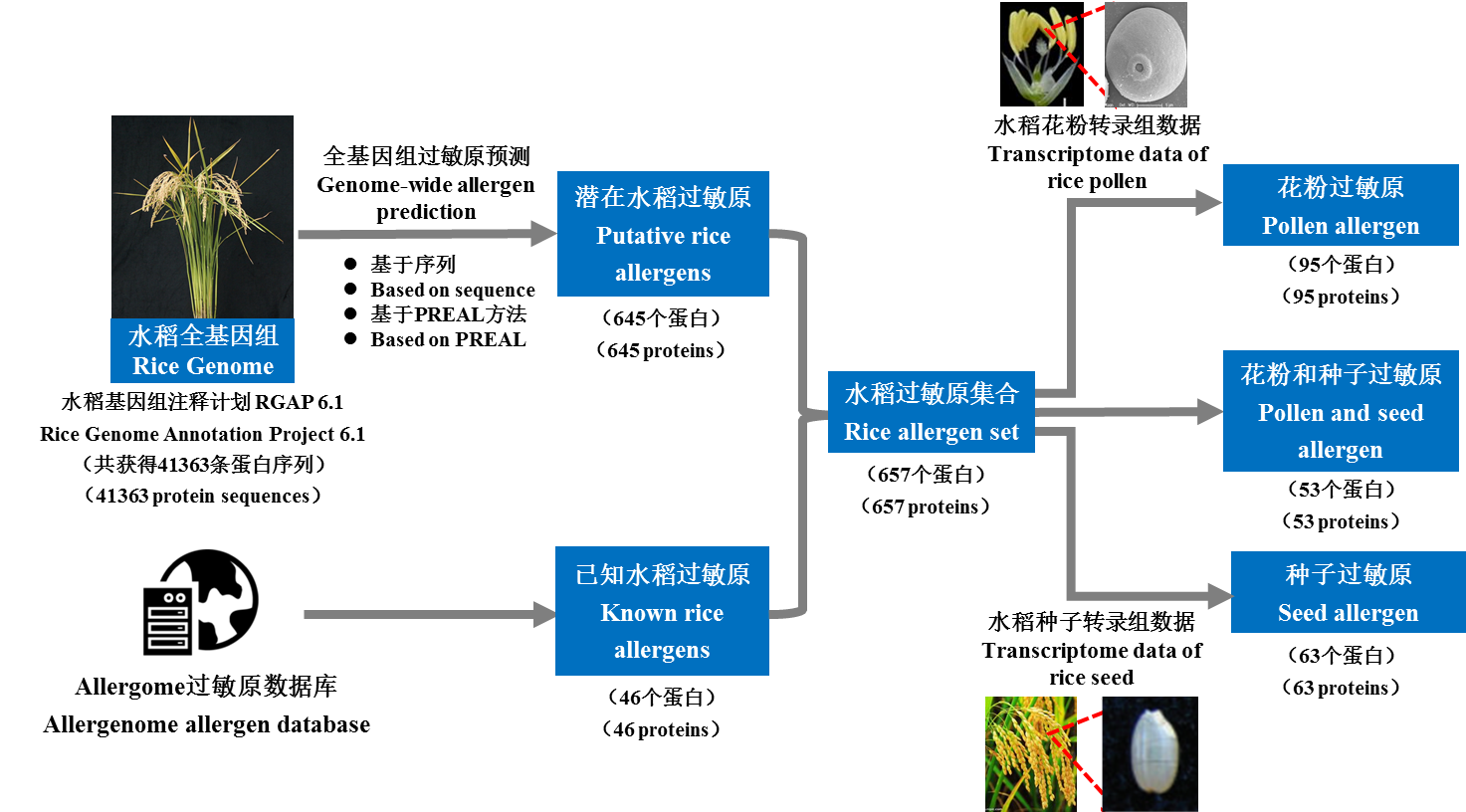


图1 水稻过敏原预测和收集流程示意图

Figure 1 Flow diagram of collection and prediction of rice allergens

**2 水稻过敏原的表达模式、蛋白家族分类和功能预测**

为了进一步了解水稻过敏原基因/蛋白的表达模式，我们对539个有探针的基因，进行了在水稻9种组织时期表达情况的聚类分析；结果表明：水稻过敏原基因/蛋白在植物体内有着多样的表达模式（图2A），其中可以明显的观察到一些过敏原在水稻多个组织中高表达（图2A，紫色基因簇I），而另一些过敏原则较特异性地在成熟花粉中高表达（图2A，橙色基因簇II）。水稻过敏原基因/蛋白表达模式的不同，预示着过敏原可能在植物体内的不同部位或不同时期，参与着不同的生物学过程。

我们进一步对水稻过敏原的功能进行GO功能预测分析，结果表明，水稻过敏原参与广泛的细胞代谢过程（图2B）；其中约40%过敏原富集在代谢过程（GO：0008152）和细胞过程（GO：0009987）中；代谢过程主要是初级代谢过程（primary metabolic process），主要为一些酶蛋白和结合功能蛋白，包括半胱氨酸型肽酶（cysteine-type peptidase）、异构酶（isomerase）、氧化还原酶（oxidoreductase）、细胞骨架（cytoskeleton）相关蛋白等；此外，一些过敏原蛋白也参与刺激响应（GO：0050896）和生物调控过程（GO：0065007）。

为了进一步了解水稻过敏原的表达特征并分析其功能，我们对这些潜在过敏原进行了蛋白家族的分类。根据Radauer和Breiteneder对已鉴定过敏原的研究，种子植物中存在2500多个蛋白质家族，而仅有59个蛋白质家族包含过敏原，这表明过敏原在家族分布上有其一定的特征(Radauer and Breiteneder, 2006)。在我们的研究中发现的657个过敏原分布到了总共159个不同的蛋白家族中，包括109个已知家族和50个未知家族（图2C）。这些蛋白家族包括了其他常见的过敏原家族如扩张蛋白家族（Expansin）、前纤维蛋白家族（Profilin）、Bet v 1过敏原家族和Cupin超家族等。其中所含过敏原最多的家族为肽酶家族（Peptidase），其次为热激蛋白家族（Heat shock protein，HSP）以及水解酶家族（Glycoside hydrolase），这三个家族也均有已知过敏原报道。在水稻过敏原蛋白之中，还有一些蛋白（31个蛋白）并未得到明确的蛋白家族分类。水稻过敏原的蛋白家族分布暗示，过敏原可能来自于一些大的蛋白家族中，而这些相似的家族成员可能在功能分化的过程中产生了致敏性的变化。

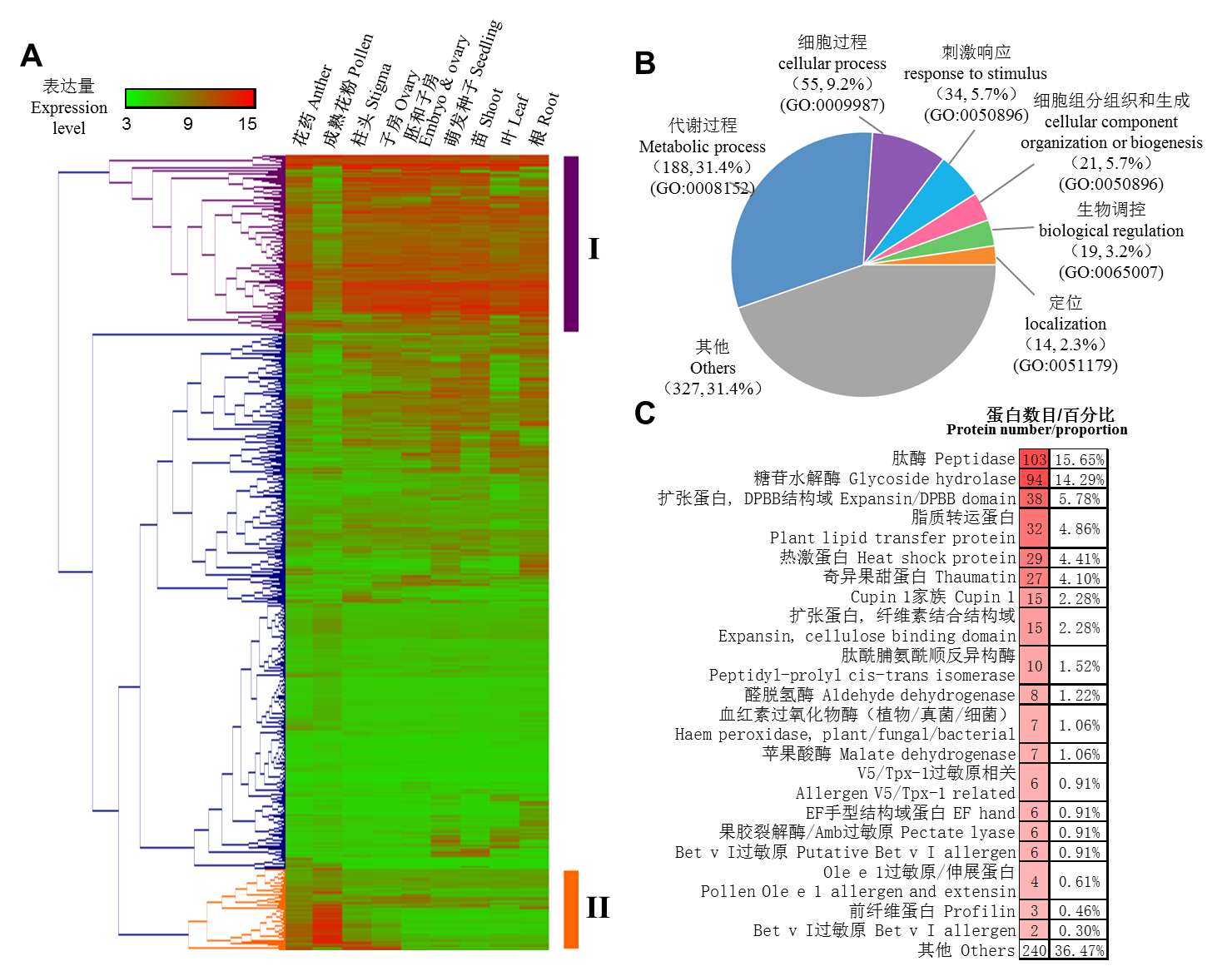


图2 水稻过敏原的表达模式分析、功能预测和蛋白家族分类

Figure 2 Expression pattern, functional prediction and protein family classification of rice allergens

A: 水稻过敏原在9种不同组织部位的表达模式和聚类分析；B：水稻过敏原的功能分类；C：水稻过敏原的蛋白家族分类统计。

**3 水稻过敏原同源蛋白的物种分布**

为了进一步了解水稻过敏原在植物中的分布及进化情况，我们以水稻过敏原蛋白序列为检索序列，通过BLASTP最佳匹配找到在25个从低等到高等的绿色植物中的同源序列（包括种内同源蛋白），按之前划分的蛋白家族分析这些过敏原同源蛋白的物种分布情况。结果表明，从绿藻到被子植物找到的同源蛋白逐渐增多，在和水稻亲缘关系单子叶植物中有较多的同源蛋白（图3）。我们已经知道在被子植物进化过程中，其祖先经历了多次全基因组复制事件(Adams and Wendel, 2005; Bowers et al., 2003)，在某些物种分化之后，也有基因自身的复制事件，因此，我们推测水稻过敏原家族也可能经历过成员的增多和分化，这可能与全基因组复制事件或独立复制事件相关。

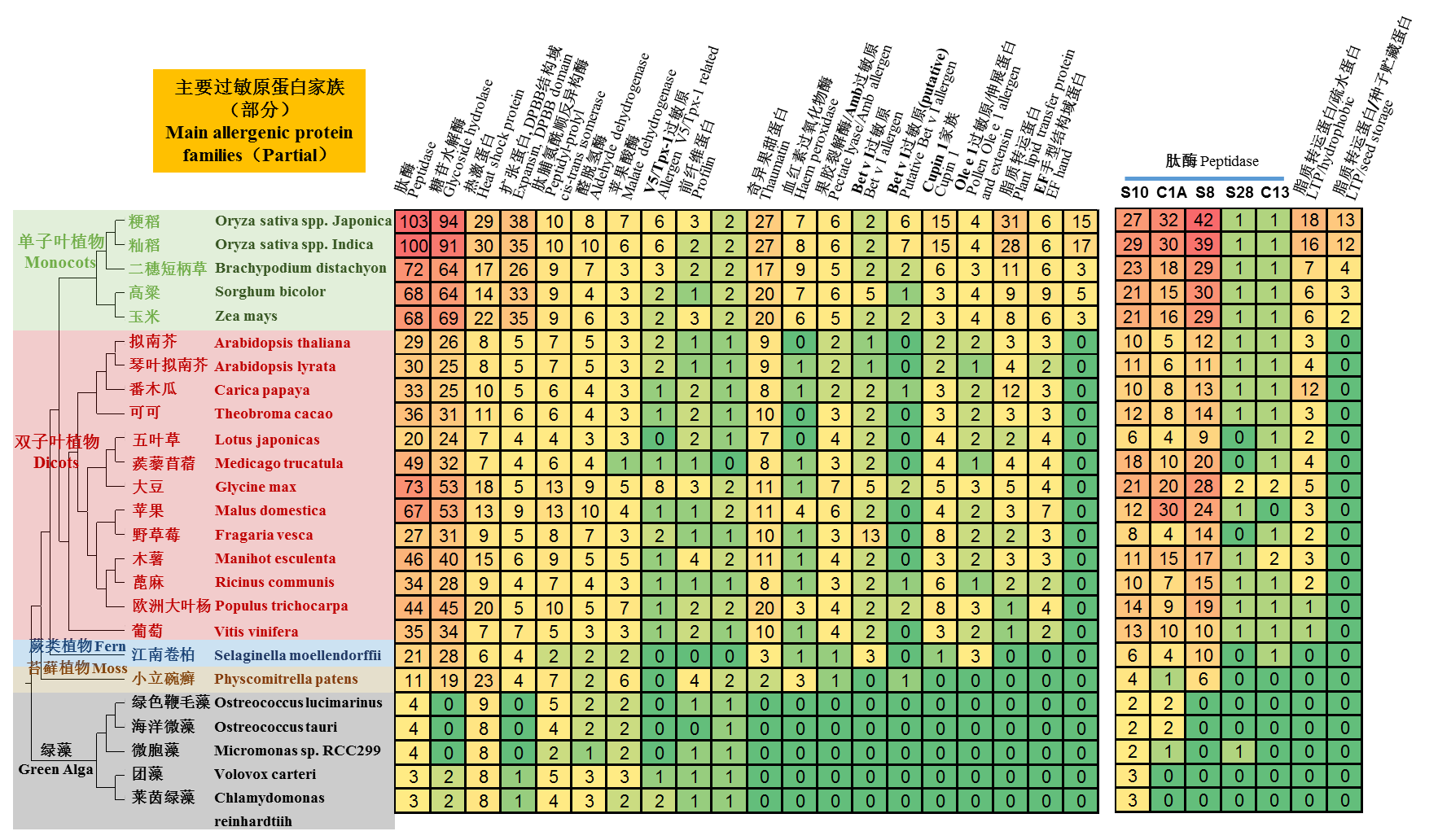


图 3水稻过敏原同源蛋白的物种分布

Figure 3 Taxonomic distribution of rice allergen homologues

另外不同家族水稻过敏原同源蛋白在单子叶植物中的数目差异较小，这也可能暗示着这些过敏原在禾本科植物中均有着广泛且相似的分布情况。而在双子叶植物中，过敏原家族同源蛋白的数目已经开始发生较大的变化，一些家族如肽酶和糖苷水解酶在双子叶植物中也有着大量的同源蛋白，而数目没有单子叶植物中多。值得注意的是，在我们分析的双子叶物种中，大豆含有最多的肽酶和糖苷水解酶同源蛋白，而大豆也是日常生活中最重要的食物过敏原之一。物种中过敏原同源蛋白的数目可能在一定程度上暗示着该物种致敏的危险性。

我们的研究结果还发现，一些过敏原家族：如热激蛋白家族、前纤维原蛋白家族等的致敏蛋白似乎有着古老的起源，水稻中这些家族中的致敏性蛋白均能在低等植物中找到非常相似的同源基因（图3）。而另一些致敏家族中的致敏蛋白只在高等植物（单、双子叶植物）中匹配到同源基因，表明其起源较晚或在进化过程中发生了较大变异。例如扩张蛋白家族中的一些成员（包含纤维素结合域）仅在单子叶植物中匹配到同源基因，这些成员可能在禾本科植物进化的过程中累积序列变化承担了新的功能。在一些大的致敏蛋白家族中，不同的亚家族的同源蛋白也有着不同的物种分布情况，如肽酶的不同亚家族，S10、1A和S8亚家族中过敏原成员数目较多，但S8亚家族在低等生物(绿藻、苔藓植物)中并不存在同源蛋白（图3）。

**4 水稻花粉过敏原和种子过敏原的表达模式分析**

我们已经知道，花粉过敏原和种子过敏原是最重要两种植物过敏原。因此，我们进一步系统分析了水稻中的花粉过敏原和种子过敏原的情况。根据过敏原是否在水稻花粉和种子转录芯片中表达情况和特异性，水稻中在花粉、种子中表达的过敏原基因被归类为花粉过敏原（95个）、种子过敏原（64个）以及花粉和种子共有过敏原（53个）（图1）。

通过比较这些过敏原在水稻12个不同组织的表达情况，我们可以看到这三种类型的过敏原在表达模式上有着较大的差异。花粉过敏原多在花药、花粉发育后期（双核、三核花粉和成熟花粉）中高表达（图4A），种子过敏原多在胚、愈伤组织中表达（图4B），而在花粉和种子均筛选到的过敏原大多是全局表达的基因（图4C）。

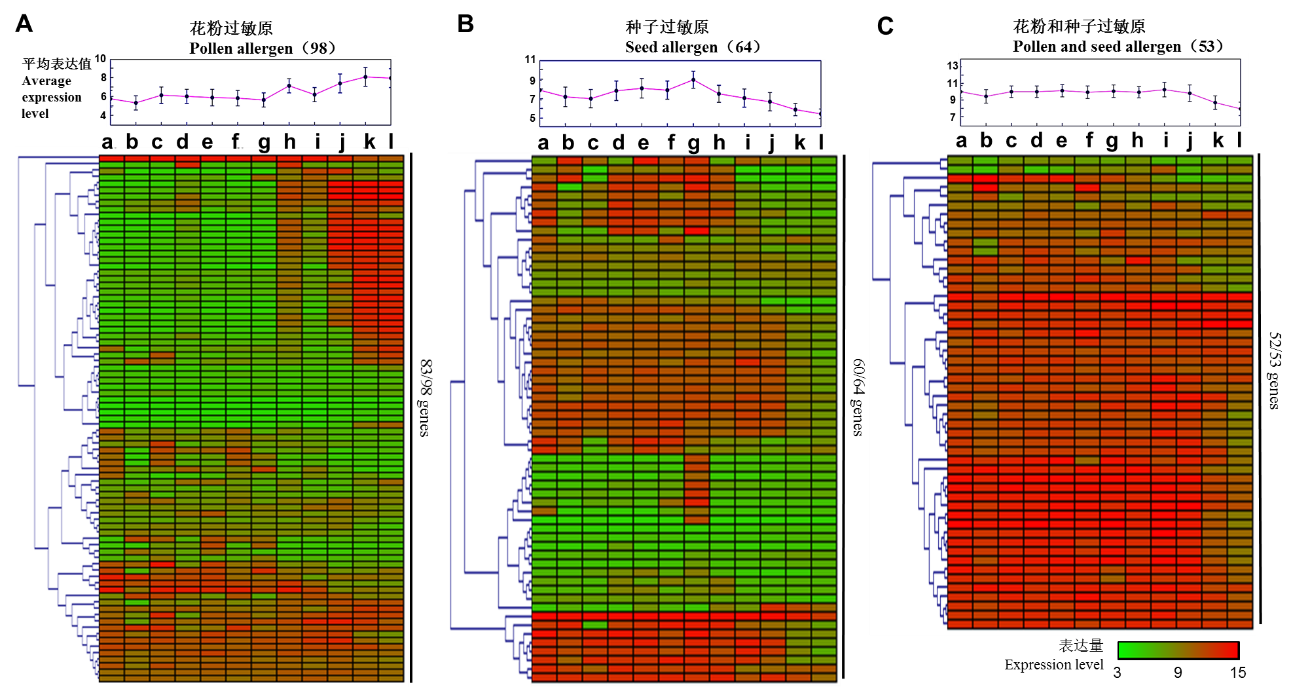


图 4水稻花粉和种子过敏原的表达模式比较

Figure 4 Comparison of expression patterns of rice pollen and seed allergens

A：水稻花粉过敏原的表达模式分析；B：水稻种子过敏原的表达模式分析，C：花粉和种子共有过敏原的表达模式分析（表达组织部位对应关系，a：苗，b：叶片，c：根，d：子房，e：胚和子房，f：萌发种子，g：愈伤组织，h：花药，i：单核花粉，j：双核花粉，k：三核花粉，l：成熟花粉）

**5 水稻花粉过敏原和种子过敏原的蛋白家族和功能分析**

花粉和种子中共有的蛋白家族多为一些行使更加普遍功能的蛋白家族，如参与胁迫响应的热激蛋白和过氧化物歧化酶等，也包括一些参与代谢过程的催化酶类，如肽酶、苹果酸酶和烯醇酶等。这也符合它们广泛表达的特征。花粉、种子过敏原的蛋白家族分布还有着一些差异，花粉过敏原和种子过敏原都有着一些独特的蛋白家族。如种子中含有大量的脂质转运蛋白（LTP）/贮藏蛋白家族成员和淀粉合成酶，而花粉中的特有的过敏原则包括一类扩张蛋白、果胶裂解酶以及Bet v I家族等。花粉和种子中也共有一类脂质转运蛋白，但种子中的转运蛋白的数目占绝大优势。

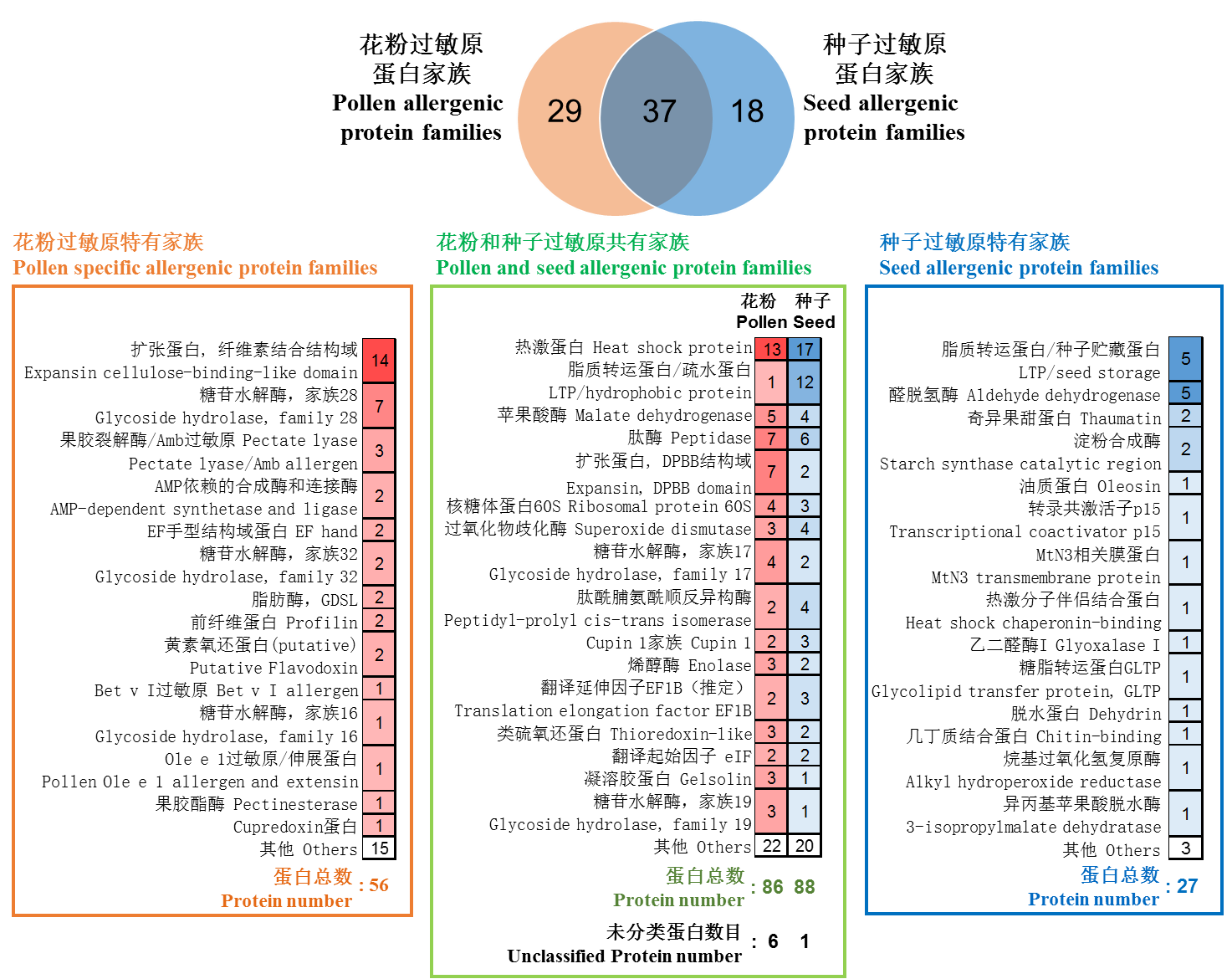


图 5 水稻花粉、种子过敏原的蛋白家族统计

Figure 5 Summary of pollen and seed allergens protein families in rice

进一步的功能富集分析表明，花粉和种子过敏原在细胞组分、分子功能和生物学过程中的差异并不明显。这也可能是因为过敏原的数目过少，在一些GO条目中没有得到富集。花粉过敏原和种子过敏原在细胞组分上主要为细胞部分（GO:0044464），大分子复合物（GO:0032991）等；在分子功能上主要为结合功能（GO:0005488）和催化功能（GO:0003824）；在生物学功能上主要为细胞过程（GO:0009987）和代谢过程（GO:0008152）。在生物学过程中，花粉过敏原参与生殖过程（GO：0000003），仅有花粉过敏原中富集到肌动蛋白纤维过程（GO:0030029）和微管过程（GO:0007017），这可能与花粉萌发过程中的功能相关。另外值得注意的是，与花粉过敏原相比，种子过敏原更多地属于胞外区域（GO:0005576），且参与外部封装结构（GO:0030312），这可能与花粉外壁相关。种子过敏原则更多地参与刺激响应（GO:0050896）、定位（GO:0051179）和转运过程（GO:0006810）。

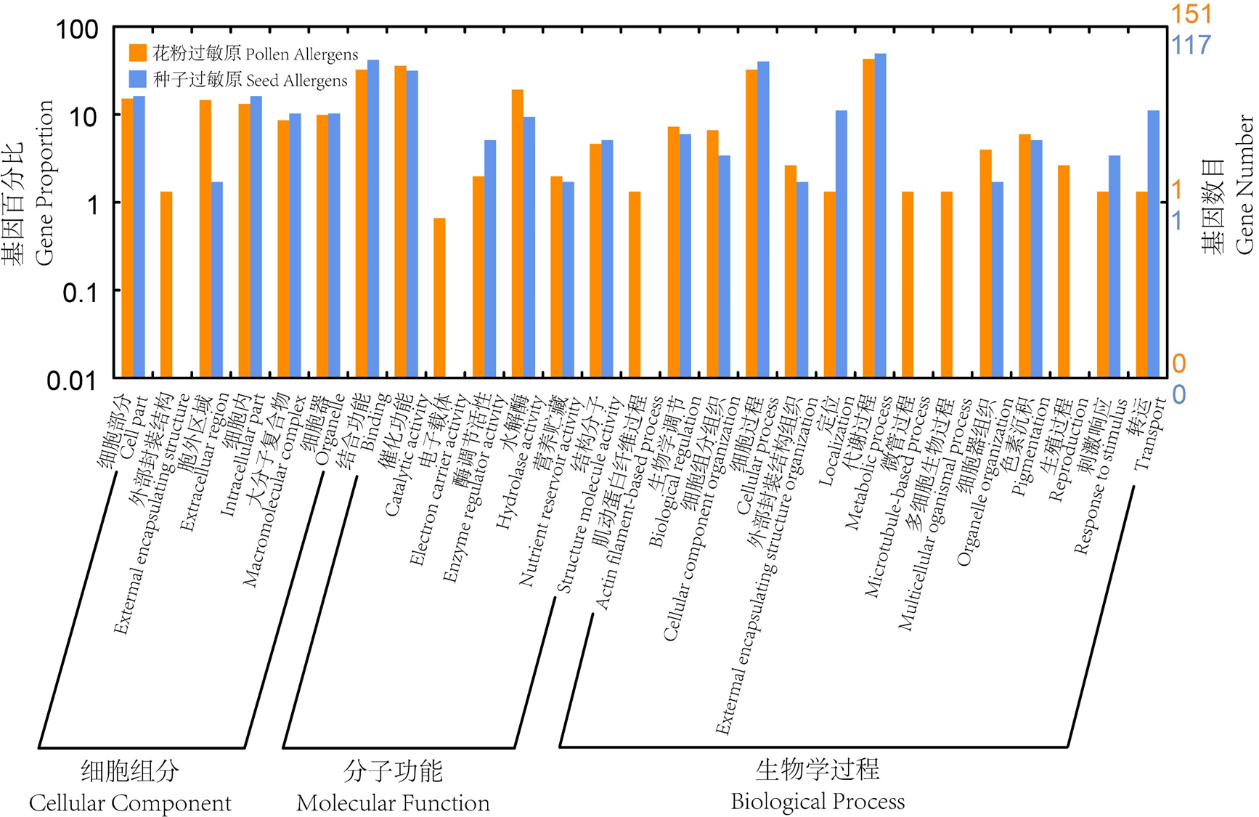


图 6 水稻花粉、种子过敏原的蛋白家族统计

Figure 6 Summary of pollen and seed allergens protein families in rice

**2讨论**

水稻是重要的粮食作物和模式植物，本研究的研究全面的预测了水稻中的潜在过敏原蛋白，并筛选出了高风险的花粉过敏原和种子过敏源蛋白。我们首先利用两种生物信息学方法在全基因组层面上预测了水稻过敏原，两种生物信息学预测方法同时使用保证了较高的预测准确性，检测到了74%（34/46）的已知过敏原。仍有一些已知过敏原未同时被两种方法预测到，表明其准确性仍有提升空间。通过对过敏原表达模式分析，表明水稻过敏原基因的表达模式多样，而表达模式和表达量与植物哪些部位导致人类过敏有着直接关系(Songnuan, 2013)。如普遍高表达的过敏原基因则有更多机会被人类接触而导致过敏；而花粉特异表达的过敏原基因，则需注意防范对花粉的接触；另外，预测到的一些普遍低表达的基因，虽然其表达的蛋白可能有着过敏原的序列和结构特征，但其表达量较低也意味着其导致过敏的可能性较小。未来在过敏原的预测方法中加入对蛋白表达模式的评估可能有助于更准确的预测蛋白实际的致敏可能性。对过敏原进行表达模式研究，也有助于帮助我们找到植物具有致敏危险组织。

通过分析水稻过敏原的同源蛋白在25种植物物种中的分布情况，我们对植物过敏原的物种分布有了一个全面直观的了解。不同物种中水稻过敏原同源蛋白的分布情况，能帮助我们快速找到特定物种中的潜在过敏原，特别是未经测序的物种。同源蛋白在低等到高等植物中的分布模式暗示了其不同的进化历程。由于在与人类的长期接触中，过于保守的过敏原可能会导致致敏性的减弱或消除。所以我们推测过敏性的产生也许并非完全是偶然，一些蛋白家族中的过敏原可能随着功能的分化而产生，而其致敏性随着新功能被固定下来。如禾本科植物中的一支β-扩张蛋白具有致敏性（水稻中为Ory s 1），是重要的花粉过敏原，而双子叶植物中并未观察到这类β-扩张蛋白。研究表明，这类β-扩张蛋白可能通过进化以更好的适应禾本科花粉中的特殊底物——高取代的木聚糖(Sampedro et al., 2014)。不幸的是，这一功能的分化可能也产生了致敏性，产生了group-1禾本科花粉过敏原(Flicker et al., 2006)。

此外，由水稻花粉和种子转录组数据，本研究筛选出了花粉过敏原和种子过敏原。花粉过敏原和种子过敏原在表达模式上均有时空偏向性，花粉过敏原倾向于在花粉发育后期和成熟花粉中高表达，而种子过敏原倾向于在种子、胚和子房以及愈伤组织中高表达。同时基因家族和功能的分布表明，不仅是表达模式上有差异，这些花粉和种子过敏原有着一些特有的基因家族。种子中含有大量的脂质转运蛋白（LTP）/贮藏蛋白家族成员，这是水稻种子中一种重要的过敏原，已知的Ory s 14过敏原均属于这一家族。另外，淀粉合成酶也是种子中特有的过敏原之一，包括已知的Ory s GBSS I过敏原。花粉中的特有的过敏原则包括一类扩张蛋白、果胶裂解酶。结合功能分析，暗示着过敏原在植物生理发育过程中发挥着重要作用。这些花粉或种子特有的过敏原与其特定的功能相关，如种子中需要大量淀粉合成酶和贮藏蛋白行使功能保证种子发育和功能，而花粉壁的独特结构则需要扩张蛋白和果胶裂解酶来保证花粉的萌发过程(Songnuan, 2013; Huang et al., 2015)。这暗示了植物不同组织部位的过敏原与该组织部位特定的发育、代谢过程有着一定的关联，这种符合组织部位需求的功能分化可能与致敏性的产生相关。

本研究通过对水稻过敏原蛋白表达模式、蛋白家族分布以及同源蛋白的物种分布的分析能够帮助预测其他物种中的过敏原，从而帮助预防和治疗花粉过敏症。同时，我们推测过敏原在植物中承担着重要的生物学功能，了解这些过敏原的生物学功能有助于我们理解致敏性产生的机理。

**3材料和方法**

**1.1 水稻基因组和转录组数据收集**

水稻基因组数据来自水稻基因组注释计划（Rice Genome Annotation Project，RGAP6.1，[http://rice.plantbiology.msu.edu/），](http://rice.plantbiology.msu.edu/），RGAP)主要包含41363条蛋白序列。

水稻转录组芯片数据均来自于NCBI GEO (Gene Expression Omnibus)和EMBL-EBI ArrayExpress (http://www.ebi.ac.uk/arrayexpress/)数据库 (Barrett et al., 2009； Kolesnikov et al., 2015)，其中水稻花粉转录组数据包括4组芯片数据，GEO编号如下：GSM433634、 GSM433635、GSM433636和GSM433637；水稻种子则包括3组数据，来自ArrayExpress数据集，编号E-MEXP-1766。所有芯片数据均来自Affymetrix芯片平台。

水稻不同组织/时期的芯片表达数据获取自Rice Oligonucleotide Array Database (ROAD)数据库(<http://www.ricearray.org/>) (Cao et al., 2012)。水稻中共选取了12组不同时期和组织的表达数据，包括四组花粉发育时期：单核花粉期（uni-nucleated pollen）、双核花粉期（bi-cellular pollen）、三核花粉期（tri-cellular pollen）和成熟花粉期（mature pollen），以及花药（Anther）、子房（Ovary）、子房和胚（embryo and ovary）、萌发种子（Seeding）、愈伤组织（Callus）、根（Root）、苗（Shoot）和叶片（Leaf）。

**1.2****花粉过敏原的预测**

为了保证花粉过敏原预测的准确性，我们在Wang等的过敏原的对比分析基础上(Wang et al., 2013)，同时应用基于序列的分析方法和最大相关最小冗余的特征选择方法预测过敏原，被两种方法均预测为过敏原的蛋白进入到后续分析流程。如前所述，基于序列的预测方法由FAO/WHO提出(FAO/WHO, 2003)，这一方法基于被预测蛋白与已知蛋白氨基酸序列之间的相似性进行致敏性预测。连续8个氨基酸与已知过敏原完全匹配则预测为花粉过敏原（rule 1），通过proAP（http://gmobl.sjtu.edu.cn/proAP/main.html）上传蛋白质序列文件进行预测。使用PREAL方法进行最大相关最小冗余的特征选择预测，使用所有过敏原构建预测模型，概率阈值为默认参数0.5，在服务器上使用PERL脚本进行过敏原预测(Wang et al., 2013)。

**1.3** **花粉和种子表达基因筛选**

我们首先通过MAS 5.0算法对水稻成熟花粉和种子的芯片表达值进行标准化，并用该算法的A/P call判断每个探针是否检测到表达信号。筛选出在一半以上实验重复中都检测到信号（A/P call为present）的探针作为候选探针。通过水稻Affymetrix平台对应的CDF文件对芯片探针进行注释，将候选探针对应转换为水稻基因ID。

**1.4** **表达谱分析和聚类**

为了消除不同实验造成的批次效应（batch effects），我们使用ComBat算法包(Chen et al., 2011; Johnson et al., 2007)，通过经验贝叶斯模型矫正表达值，以矫正后的表达值作为最终表达值。然后，通过Genesis软件进行表达谱的可视化分析，对原始表达值进行对数转换取log2值，并进行作图。聚类分析选择层次聚类方法（hierarchical clustering），聚类规则采用平均距离法（Average linkage clustering）进行基因表达谱聚类分析(Sturn et al., 2002)。

**1.5 水稻基因注释和基因功能富集的统计分析**

通过水稻过敏原的基因信息，使用PLAZA数据库版本2.5 (<http://bioinformatics.psb.ugent.be/plaza/>) (Van Bel et al., 2012) 以及PANTHER基因功能分类系统 (http://pantherdb.org/) (Mi et al., 2013)进行功能预测，GO注释及富集分析，保留显著富集的GO条目（P ≤ 0.05）。GO比较分析结果由WEGO作图展示(Ye et al., 2006)。

**1.6 花水稻过敏原蛋白家族分类和物种分布分析**

水稻过敏原基因/蛋白根据Pfam蛋白家族数据库（版本27.0）(<http://pfam.xfam.org/>) (Finn et al., 2014)以及植物基因家族数据库(http://green.dna.affrc.go.jp/PGF-DB/)分类到各个蛋白家族，从PLAZA数据库中选取25个物种，进行物种分布分析，它们从低等植物到高等植物均有分布。

这些物种具体包括5种绿色藻类：绿色鞭毛藻(*Ostreococcus lucimarinus*)、海洋微藻(*Ostreococcus tauri*)、微胞藻RCC299(*Micromonas* *sp.* RCC299)、团藻（*Volvox carteri*）、莱茵绿藻（*Chlamydomonas reinhardtii*）；1种苔藓植物：小立碗藓（*Physcomitrella patens*）；1种蕨类植物：江南卷柏（*Selaginella moellendorffii*）；18种被子植物（Angiosperms）包括5种典型单子叶植物和13种双子叶植物，单子叶植物包括玉米（*Zea mays*）、高粱（*Sorghum bicolor*）、二穗短柄草（*Brachypodium distachyon*）、2种水稻亚种：粳稻（*Oryza sativa ssp. Japonica*）、籼稻（*Oryza sativa ssp. Indica*）；双子叶植物包括葡萄（*Vitis vinifera*）、可可（*Theobroma cacao*）、番木瓜（*Carica papaya*）、拟南芥（*Arabidopsis thaliana*）、琴叶拟南芥（*Arabidopsis lyrata*）、欧洲大叶杨（*Polulus trichocarpa*）、蓖麻（*Ricinus communis*）、木薯（*Manihot esculenta*）、草莓（*Fragaria vesca*）、苹果（*Malus domestica*）、大豆（*Glycine max*）、蒺藜苜蓿（*Manihot esculenta*）、五叶草（*Lotus japonicus*）。

通过BLASTP(e-value<10-5)找到检索蛋白在所比较的物种中匹配到的最相似蛋白作为同源蛋白，若最优匹配的蛋白在物种自身中有比最优匹配bit socres更高的蛋白，则将这些蛋白认作种内同源蛋白，也一同作为检索蛋白的同源蛋白。

**作者贡献**

陈淼林负责具体分析工作，数据收集整理及文章撰写；许杰为通讯作者，负责研究思路指导，文章修改和审定。全体作者都阅读并同意最终文本。

**致谢**

本研究由国家自然科学基金面上项目(31370026、31570312和31500297)资助。

**参考文献**

Adams, K.L., Wendel, J.F., 2005. Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol., 8(2): 135–41.

http://www.sciencedirect.com/science/article/pii/S1369526605000051

Bowers, J.E., Chapman, B.A., Rong, J.K., Paterson, A.H., 2003. Unravelling Angiosperm Genome Evolution by Phylogenetic Analysis of Chromosomal Duplication Events. Nature, 422(6930): 433-438.

http://dx.doi.org/10.1038/nature01521

Cao, P., Jung, K.-H., Choi, D., Hwang, D., Zhu, J., Ronald, P.C., 2012. The rice oligonucleotide array database: an atlas of rice gene expression. Rice, 5(1): 17-17.

http://link.springer.com/article/10.1186%2F1939-8433-5-17

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., Liu, C., 2011. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One, 6(2): e17238.

http://europepmc.org/articles/PMC3046121/

Cui, J., Han, L.Y., Li, H., Ung, C.Y., Tang, Z.Q., Zheng, C.J., Cao, Z.W., Chen, Y.Z., 2007. Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. Mol. Immunol., 44(4): 514–520.

http://www.sciencedirect.com/science/article/pii/S0161589006000368

Enrique, E., Ahrazem, O., Bartra, J., Latorre, M.D., Castelló, J.V., Mateo, J.A. De, Montoya, E., Malek, T., Barber, D., Salcedo, G., 2005. Lipid transfer protein is involved in rhinoconjunctivitis and asthma produced by rice inhalation. J. Allergy Clin. Immunol., 116(4): 926–928.

http://www.sciencedirect.com/science/article/pii/S0091674905013709

FAO/WHO, 2003. Report of a Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Bio-technology Rome.

http://www.fao.org/ag/agn/food/pdf/bi03al.pdf

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., Punta, M., 2014. Pfam: The protein families database. Nucleic Acids Res., 42(Database issue): D222-30.

http://nar.oxfordjournals.org/content/42/D1/D222.full

Flicker, S., Steinberger, P., Ball, T., Krauth, M.T., Verdino, P., Valent, P., Almo, S., Valenta, R., 2006. Spatial clustering of the IgE epitopes on the major timothy grass pollen allergen Phl p 1: Importance for allergenic activity. J. Allergy Clin. Immunol., 117(117): 1336–1343.

http://www.sciencedirect.com/science/article/pii/S0091674906003241

Gendeh, B.S., Murad, S., Razi, A.M., Abdullah, N., Mohamed, A.S., Kadir, K.A., 2000. Skin prick test reactivity to foods in adult Malaysians with rhinitis. Otolaryngol. Head Neck Surg., 122(5):758-762.

http://www.sciencedirect.com/science/article/pii/S0194599800702117

Goff, S.A., Briggs, S., 2002. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). Science, 296(5565): 92-100.

http://classic.sciencemag.org/content/296/5565/92.long

Hirano, K., Hino, S., Oshima, K., Okajima, T., Nadano, D., Urisu, A., Takaiwa, F., Matsuda, T., 2013. Allergenic potential of rice-pollen proteins: expression, immuno-cross reactivity and IgE-binding. J. Biochem., 154(2): 195-205.

http://jb.oxfordjournals.org/content/154/2/195.short

Huang Y., Zhu L., Luo R., Zhang D., Xu J., 2015. Evolutionary and Expression Profile Analysis of the Non-specific Lipid Transfer Proteins Genes Superfamily in and Rice. Genomics and Applied Biology, 34(7): 1510-1521. (黄岳, 朱璐, 罗人和, 张大兵, 许杰, 2015. 拟南芥和水稻非特异脂质转运蛋白的分子进化、表达模式以及在花药发育中的功能分析. 基因组学与应用生物学, 34(7): 1510-1521.)

http://xueshu.baidu.com/s?wd=paperuri%3A%28874a447131d15b5b58fb8736db08b9f4%29&filter=sc\_long\_sign&tn=SE\_xueshusource\_2kduw22v&sc\_vurl=http%3A%2F%2Fwww.cnki.com.cn%2FArticle%2FCJFDTotal-GXNB201507037.htm&ie=utf-8

Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1): 118–27.

http://biostatistics.oxfordjournals.org/content/8/1/118.full

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., Megy, K., Pilicheva, E., Rustici, G., Tikhonov, A., Parkinson, H., Petryszak, R., Sarkans, U., Brazma, A., 2015. ArrayExpress update--simplifying data submissions. Nucleic Acids Res., 43(Database issue): D1113-6.

http://nar.oxfordjournals.org/content/43/D1/D1113

Kosugi ., T., Saitoh ., S., Tamaki ., N., Shimoji ., K., Kakazu ., T., Saitoh ., A., Ijyu ., M., Agata ., H., 1992. Evaluation of the sensitized condition of patients with allergic diseases in Okinawa using the MAST allergy system. Arerugi, 41(7): 766–771.

http://europepmc.org/abstract/MED/1444829

Larché, M., Akdis, C. a, Valenta, R., 2006. Immunological mechanisms of allergen-specific immunotherapy. Nat. Rev. Immunol., 6(6):725–732.

http://linkinghub.elsevier.com/retrieve/pii/S0091-6749(10)01850-6

Mi, H., Muruganujan, A., Thomas, P.D., 2013. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res., 41(Database issue): D377-D386.

http://nar.oxfordjournals.org/content/41/d1/d377

Pawankar, R., Canonica, G.W., ST Holgate, S.T., Lockey, R.F., Blaiss, M., 2013. The WAO White Book on Allergy (Update. 2013). WAO ed. Milwaukee, Wisconsin World Allergy Organ.

http://www.worldallergy.org/UserFiles/file/WhiteBook2-2013-v8.pdf

Radauer, C., Breiteneder, H., 2006. Pollen allergens are restricted to few protein families and show distinct patterns of species distribution. J. Allergy Clin. Immunol., 117(1): 141–7.

http://www.sciencedirect.com/science/article/pii/S0091674905020580

Saha, S., Raghava, G.P.S., 2006. AlgPred: Prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res., 34(Web Server issue): W202–W209.

http://nar.oxfordjournals.org/content/34/suppl\_2/W202

Sampedro, J., Guttman, M., Li, L., Cosgrove, D.J., 2014. Evolutionary divergence of β-expansin structure and function in grasses parallels emergence of distinctive primary cell wall traits. Plant J. 81(1): 108–120.

http://onlinelibrary.wiley.com/doi/10.1111/tpj.12715/abstract

Songnuan, W., 2013. Wind-pollination and the roles of pollen allergenic proteins. Asian Pac. J. Allergy Immunol., 31(4): 271–6.

http://europepmc.org/abstract/med/24383968

Stadler, M.B., Stadler, B.M., 2003. Allergenicity prediction by protein sequence. FASEB J., 17(9): 1141–3.

http://www.fasebj.org/content/early/2003/06/02/fj.02-1052fje.long

Sturn, A., Quackenbush, J., Trajanoski, Z., 2002. Genesis: cluster analysis of microarray data. Bioinformatics, 18(1): 207–208.

http://bioinformatics.oxfordjournals.org/content/18/1/207

Takumi, K., Udaka, J., Kimoto, M., Koga, T., Tsuji, H., 2000. Structural and immunochemical homologies between foxtail millet glutelin 60 kDa and starch granule-bound starch synthase proteins from rice, barley, corn and wheat grains. J. Nutr. Sci. Vitaminol., 46(2): 109–112.

http://xueshu.baidu.com/s?wd=paperuri%3A%2852c7102c24478b4f8b80a6eb45d48f0c%29&filter=sc\_long\_sign&tn=SE\_xueshusource\_2kduw22v&sc\_vurl=http%3A%2F%2Fwww.ncbi.nlm.nih.gov%2Fpubmed%2F10885800&ie=utf-8

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., Vandepoele, K., 2012. Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform. PLANT Physiol, 158(2): 590-600.

http://www.jstor.org/stable/41435393?seq=1#page\_scan\_tab\_contents

Wang, J., Zhang, D., Li, J., 2013. PREAL: prediction of allergenic protein by maximum Relevance Minimum Redundancy (mRMR) feature selection. BMC Syst. Biol. 7 Suppl 5, S9.

http://link.springer.com/article/10.1186%2F1752-0509-7-S5-S9

Xu, H., Theerakulpisut, P., Goulding, N., Suphioglu, C., Singh, M.B., Bhalla, P.L., 1995. Cloning, expression and immunological characterization of Ory s 1, the major allergen of rice pollen. Gene, 164(2): 255–259.

http://www.sciencedirect.com/science/article/pii/037811199500527D

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., Wang, J., Li, S., Li, R., Bolund, L., 2006. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res., 34(2) : W293–W297.

http://nar.oxfordjournals.org/content/34/suppl\_2/W293.full

Yang, Q., Ting, O., Yan, H., Yang, J., Li, L., Xia, X., 2015. Research progress on pollen allergen. Chinese Agric. Sci. Bull. 24: 163–167. (杨琼梁, 欧阳婷, 颜红, 杨晶, 李玲, 夏新华, 2015. 花粉过敏的研究进展. 中国农学通报, 24: 163–167.)

http://xueshu.baidu.com/s?wd=paperuri%3A%280418e5ff0b8a8cbb1b4c6e3d66326100%29&filter=sc\_long\_sign&tn=SE\_xueshusource\_2kduw22v&sc\_vurl=http%3A%2F%2Fen.cnki.com.cn%2FArticle\_en%2FCJFDTotal-YXZS201003008.htm&ie=utf-8

Yang, Y., Kan, J.Q., Zhao, G.H., Shen mei, F.U., 2004. Food Allergy and Food Allergen. J. Cereal. Oils., 3: 43-45 (杨勇, 阚建全, 赵国华, 付陈梅, 2004. 食物过敏与食物过敏原. 粮食与油脂, 3: 43-45.)

http://xueshu.baidu.com/s?wd=paperuri%3A%28eaef3f94a5c55ada6b21d8250d3b1bbb%29&filter=sc\_long\_sign&tn=SE\_xueshusource\_2kduw22v&sc\_vurl=http%3A%2F%2Fen.cnki.com.cn%2Farticle\_en%2Fcjfdtotal-lsyy200403017.htm&ie=utf-8