저작자표시-비영리-변경금지 2.0 대한민국

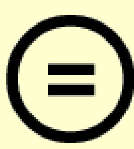
이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.


변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

농학박사학위논문

곰팡이 기능 및 비교 유전체학을 위한 유전체 탐색기와 실험 정보 관리 시스템

Genome browser and laboratory information management system for functional and comparative fungal genomics

2013년 8월

서울대학교 대학원

협동과정 농업생물공학

정 경 용

Genome browser and laboratory information management system for functional and comparative fungal genomics

A Dissertation Submitted in Partial
Fulfillment of the Requirement
for the Degree of

DOCTOR OF PHILOSOPHY

to the Faculty of
Interdiscipilinary Program in Agriculture Biotechnology

at

SEOUL NATIONAL UNIVERSITY

by

Kyongyong Jung

AUGUST, 2013

농학박사학위논문

곰팡이 기능 및 비교 유전체학을 위한 유전체 탐색기와 실험 정보 관리 시스템

지도교수 이 용 환

이 논문을 농학박사학위논문으로 제출함

2013년 5월

서울대학교 대학원
협동과정 농업생물공학

정 경 용

정 경 용의 박사학위논문을 인준함
2013년 6월

위 원 장 ________________________

부위원장 ________________________

위 원 ________________________

위 원 ________________________

위 원 ________________________

A THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Genome browser and laboratory information management system for functional and comparative fungal genomics

UNDER THE DIRECTION OF DR. YONG-HWAN LEE

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF SEOUL NATIONAL UNIVERSITY

BY
KYONGYONG JUNG

INTERDISCIPLINARY PROGRAM IN AGRICULTURE
BIOTECHNOLOGY

JUNE 2013

APPROVED AS A QUALIFIED THESIS OF KYONGYONG JUNG
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
BY THE COMMITTEE MEMBERS

CHAIRMAN ____________________________

VICE CHAIRMAN ____________________________

MEMBER ____________________________

MEMBER ____________________________

MEMBER ____________________________

Genome browser and laboratory information management system for functional and comparative fungal genomics

Kyongyong Jung

ABSTRACT

Since the full genome sequences of *Sacchromyces cerevisiae* were released in 1996, genome sequences of over 90 fungal species have become publicly available. The heterogeneous formats of genome sequences archived in different sequencing centers hampered the integration of the data for efficient and comprehensive comparative analyses. The Comparative Fungal Genomics Platform (CFGP; http://cfgp.snu.ac.kr/) was developed to archive these data via a single standardized format that can support multifaceted and integrated analyses of the data. To facilitate efficient data visualization and utilization within and across species based on the architecture of CFGP and associated databases, a new genome browser was needed.

So, we developed the SNUGB that integrates various types of genomic information derived from 91 fungal/oomycete (129 datasets) and 33 plant and animal (37 datasets)

species, graphically presents germane features and properties of each genome, and supports comparison between genomes. The SNUGB provides three different forms of the data presentation interface, including diagram, table, and text, and six different display options to support visualization and utilization of the stored information. Information for individual species can be quickly accessed via a new tool named the taxonomy browser. In addition, SNUGB offers four useful data annotation/analysis functions, including 'BLAST annotation.' The modular design of SNUGB makes its adoption to support other comparative genomic platforms easy and facilitates continuous expansion.

The SNUGB serves as a powerful platform supporting comparative and functional genomics within the fungal kingdom and also across other kingdoms. All data and functions are available at the web site http://genomebrowser.snu.ac.kr/.

As the enormous genome reveals and functional genomics becomes more advanced, large-scale phenomics presents challenges encountered in new way for the efficient handling of data. For this reason, laboratory information management system was implemented in support of managing and querying the phenotype screening data for several model organisms. Notwithstanding many efforts of applying LIMS to microbial organism research and these all systems have same disciplines but these commodities are different among the platform because each organism has different guideline for getting information from phenotype assay and there was no previously developed LIMS platform in the field of filamentous fungi which has important pathogenic fungal agents like *Magnaporthe oryzae* one of the most destructive rice

blast diseases throughout the worldwide rice cultivated area.

So, we developed LIMS for *Magnaporthe oryzae* to provide not only integrated management system for targeted knock-out but also standardized method like a guideline of phenotype assay and data acquisition formats like work process management system. In addition, we also produced close linkage that is based on locus number between phenotype data in LIMS system and genotype information in CFGP (http://cfgp.snu.ac.kr), a genome data warehouse.

This system will manage knock-out mutant by locus number, phenotype experimental assay by list of tasks, lab standard protocols, related or cited papers and influence on such research field. Sharing and comparing all standardized experimental results between all peer is supported by Global-Standard phenotype assay forms. And more easily speculate unknown genes that supported by ordered pre-existing results in LIMS database. Researchers can simply access these all system using user-friendly web-based LIMS database (http://lims.riceblast.snu.ac.kr).

LIMSMO is a valuable resource to serve the community who work on rice blast disease. Furthermore, it will shed light on a model platform for pathogenomics, not only *Magnaporthe oryzae* but also plant pathogenic fungi, to understand the disease mechanism at the systems level.

KEY WORDS: functional genomics, *Magnarporthe oryzae*, genome browser, pathogenomics, LIMS, systems biology

Student number: 2006-21477

CONTENTS

page

ABSTRACT i

CONTENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER 1. SNUGB: a versatile gnome browser supporting comparative and functional fungal genomics

page

ABSTRACT 2

INTRODUCTION 4

RESULTS 24

I. Data processing via an automated pipeline and the function of Positional Database 24

II. Modular design of SNUGB facilitates its application................................. 26

III. Properties of the fungal/oomycetes genomes archived in SNUGB 27

IV. Comparison of genome sequences of multiple isolates within species...... 32

V. Update of SNUGB. 35

VI. Functions and tools. 36

CONCLUSION AND DISCUSSION 50
LITERATURE CITED 51

CHAPTER 2. Laboratory Information Management System for the fungal pathogenomics

page

ABSTRACT 61
INTRODUCTION 62
MATERIALS AND METHODS 66
I. Development of standardized phenotype assays based on published papers 66
II. Server operation. 66
III. Database composition 66
IV. External program 67
V. Web system for LIMSMO 67
RESULTS 68
I. Development of standardized protocols. 68
II. System structure and overall function of LIMSMO 71
III. Management of experimental data via the workflow-driven experimental design of LIMSMO 80

IV. Case study. 85
V. Gold standard for phenotype assays 90
CONCLUSION AND DISCUSSION.. 94
LITERATURE CITED 97
SUPPLEMENT DATA... 102

LIST OF TABLES

CHAPTER 1

page

Table 1. List and characteristics of the genomes archived in SNUGB. 9

Table 2. Basic properties of different strains of fungal genomes deposited in SNUGB. 33

CHAPTER 2

page

Table 1. Mutant phenotype defect in transcription factor and non-transcription factor genes. 87

Supplement Table 1. Locus tagged gene and task of eight phenotype assay information in each paper 102

Supplement Table 2. Number of used material and method for phenotype assay 106

LIST OF FIGURES

CHAPTER 1

page

Figure 1. Characteristics of the 129 fungal and oomycetes genomes archived in SNUGB 30

Figure 2. Taxonomy browser 37

Figure 3. Chromosome viewer, Contig Viewer and ORF Viewer........ 38

Figure 4. Six different display methods of the genome content and properties via Graphical browser 41

Figure 5. Table and Text browser........ 44

Figure 6. BLAST annotation to catalog homelogous proteins........ 47

CHAPTER 2

page

Figure 1. The type and frequency of phenotype assays used in in published papers 70

Figure 2. The software architecture of LIMSMO 72

Figure 3. Database scheme of LIMSMO 76

Figure 4. Web based end-user interface of LIMSMO........ 77

Figure 5. Processes of data acquisition and archiving 81

Supplement Figure S1. Data acquisition of phenotype assay in published paper 121

Supplement Figure S2. Scheduled procedure to phenotype experiment assay for efficiently managed a process of assay 123

Supplement Figure S3. Front page of LIMSMO homepage 124

Supplement Figure S4. List of all locus from phenotype assay data that its public flag is published.. 125

Supplement Figure S5. Two way of view type for published data list 127

Supplement Figure S6. Published data input form... 128

Supplement Figure S7. Project management page in “My Data” section.......... 129

Supplement Figure S8. Locus of Assay task management page in My Data” section 131

Supplement Figure S9. Protocols of assay page 133

Supplement Figure S10. My protocols management page in “Protocol” section..... 135

Supplement Figure S11. Publications in LIMSMO 137

CHAPTER 1

SNUGB: a versatile genome browser supporting comparative and functional fungal genomics

ABSTRACT

Since the full genome sequences of *Sacchromyces cerevisiae* were released in 1996, genome sequences of over 90 fungal species have become publicly available. The heterogeneous formats of genome sequences archived in different sequencing centers hampered the integration of the data for efficient and comprehensive comparative analyses. The Comparative Fungal Genomics Platform (CFGP; http://cfgp.snu.ac.kr/) was developed to archive these data via a single standardized format that can support multifaceted and integrated analyses of the data. To facilitate efficient data visualization and utilization within and across species based on the architecture of CFGP and associated databases, a new genome browser was needed.

So we developed the SNUGB that integrates various types of genomic information derived from 91 fungal/oomycete (129 datasets) and 33 plant and animal (37 datasets) species, graphically presents germane features and properties of each genome, and supports comparison between genomes. The SNUGB provides three different forms of the data presentation interface, including diagram, table, and text, and six different display options to support visualization and utilization of the stored information. Information for individual species can be quickly accessed via a new tool named the taxonomy browser. In addition, SNUGB offers four useful data annotation/analysis functions, including 'BLAST annotation.' The modular design of SNUGB makes its adoption to support other comparative genomic platforms easy and facilitates continuous expansion.

The SNUGB serves as a powerful platform supporting comparative and functional genomics within the fungal kingdom and also across other kingdoms. All data and functions are available at the web site http://genomebrowser.snu.ac.kr/.

INTRODUCTION

As the number of sequenced genomes rapidly increases, search and comparison of sequence features within and between species has become an integral part of most biological inquires. To facilitate uses of the sequenced genomes, numerous bioinformatics tools have been developed; among these, genome browser plays an essential role by providing various means for viewing genome sequences and associated features (e.g., chromosomal position and context of individual genes, protein/nucleotide sequences, structures of exon/intron and promoters) via graphical and text interfaces. Widely utilized genome browsers include: (i) Ensembl (http://www.ensembl.org/), which is specialized for mammalian genomics and comparative genomics (Flicek *et al.*, 2008), (ii) UCSC Genome Browser (http://genome.ucsc.edu/), which archives genome sequences of 30 vertebrate and 24 non-vertebrate species (Kuhn *et al.*, 2007), (iii) GBrowse, a widely-used component-based genome browser (Stein *et al.*, 2002), and (iv) Map Viewer at the National Center for Biotechnology Information (NCBI), which covers a large number of organisms (Wheeler *et al.*, 2007). A new genome browser based on the Google map engine, called the X::Map Genome Browser (Yates *et al.*, 2008), contains genomes of three mammalian species and is specialized for supporting microarray analyses based on the Affymetrix platform (Okoniewski *et al.*, 2007).

Since complete *S. cerevisiae* genome sequences were released in 1996, more than 90 fungal/oomycete species have been sequenced with many additional species being

currently sequenced (Park *et al.*, 2008c). A few sequencing centers, such as the Broad Institute (http://www.broad.mit.edu/) and the JGI (http://www.jgi.doe.gov/), have sequenced most of the fungal genomes and provide their own genome browsers to support data visualization and utilization. Although they use standardized formats, such as fasta and gff3, for data presentation and distribution, each center uses its own data formats for sequences, annotation data, and other chromosomal information. In addition, some of the sequenced fungal genomes lack certain data, such as exon positions. These problems have hampered the integration and visualization of all available genome sequences via a single genome browser. As a solution for this problem, a group at Duke University (http://fungal.genome.duke.edu/) installed an open-source browser called the GBrowse (Stein *et al.*, 2002) after reannotating genome sequences of 42 fungal species from multiple sequencing centers through the use of their own annotation pipeline consisting of several gene prediction programs; large scale evolutionary analyses were conducted based on the archived genomes, demonstrating the usefulness of unified and standardized data formats (Fitzpatrick *et al.*, 2006).

A large number of sequenced fungal genomes have provided opportunities to compare genome sequences and features at multiple taxon levels, revealing potential mechanisms underpinning fungal evolution and biology (Cliften *et al.*, 2003; Cornell *et al.*, 2007; Fitzpatrick *et al.*, 2006; Galagan *et al.*, 2005; Kellis *et al.*, 2003; Soanes *et al.*, 2008; Wolfe, 2006; Xu *et al.*, 2007); however, due to the complexity and vast scale of the resulting data, presentation of these data in an easily accessible format is

challenging. To overcome this limitation, both the database construction and the pipeline/tools for comparative analyses should be carefully designed. One good example is the e-Fungi project (http://www.e-fungi.org.uk/) (Hedeler *et al.*, 2007), which archives genome sequences of 34 fungal and 2 oomycete species and supports various queries via the web interface. Comparative fungal genomics studies have been conducted using e-Fungi (Cornell *et al.*, 2007; Soanes *et al.*, 2008). Yeast Gene Order Browser (YGOD; http://wolfe.gen.tcd.ie/ygob/) (Byrne and Wolfe, 2005) archives genome sequences of the species belonging to the subphylum Sacchromycotina and provides a graphical gene order browser, which helps the dissection of evolutionary history of genome changes during yeast speciation (Scannell *et al.*, 2006). Although these platforms provide useful tools and data, only certain fungal genomes are covered, and graphical presentation of data is limited.

The Comparative Fungal Genomics Platform (CFGP; http://cfgp.snu.ac.kr/) was established to archive all publicly available genome sequences using a unified data format and to support multifaceted analyses of the stored data via a newly developed user interface named as Data-driven User Interface (Park *et al.*, 2008c). Currently, CFGP archives genome sequences of 87 fungal and 4 oomycete species (129 different datasets) and also carries genome sequences of 55 plant, animal and bacterial species (56 datasets). Taking advantage of the data warehouse and functionalities in CFGP, several databases specialized for certain gene families or functional groups have been constructed, one of which is the Fungal Transcription Factor Database (FTFD; http://ftfd.snu.ac.kr/) (Park *et al.*, 2008b). This database identified and classified all

fungal transcription factors and provides a phylogenomic platform supporting analyses of individual transcription factor families (Park *et al.*, 2006). In addition, Fungal Cytochrome P450 Database (FCPD; http://p450.riceblast.snu.ac.kr/) (Park *et al.*, 2008a), Fungal Secretome Database (FSD; http://fsd.snu.ac.kr/) (Choi *et al.*, 2010), Fungal Expression Database (FED; http://fed.snu.ac.kr/) have been constructed or are currently being constructed. The CFGP was also used to manage high-throughput experimental data and link them to corresponding genes (Choi *et al.*, 2007; Jeon *et al.*, 2007) as well as to maintain the Phytophthora database (Park *et al.*, 2008d).

To support comparative genomics analyses using CFGP and offer tools for versatile data visualization, we newly developed a highly versatile genome browser named as the Seoul National University Genome Browser (SNUGB; http://genomebrowser.snu.ac.kr/). We chose to develop a new genome browser instead of adopting one of the existing browsers in part because the adoption required conversion of the data archived in CFGP into new formats, and the existing browsers do not support the integration of additional databases, such as the InterPro and customized homologous gene databases available through SNUGB. We also wanted to have a browser based on the architecture of CFGP and associated databases so that we would be able to quickly present updated contents in these resources and seamlessly integrate new tools for data processing, visualization, and/or utilization.

The SNUGB currently covers genome sequences and associated information for 87 fungal and 4 oomycete species (129 datasets), which is the largest among the

available fungal genome browser services on the web. These 87 fungal species cover four phyla and one subphylum based on a recently revised fungal taxonomy framework (Hibbett *et al.*, 2007). It also houses genome sequences of 11 plant, 18 insect, and 3 nematodes species and human genome sequences (37 datasets), to support comparison of fungal genomes with those in other kingdoms (Table 1). The SNUGB provides lists of putative orthologous genes of all fungal ORFs and a tool for

Table 1. List and characteristics of the genomes archived in SNUGB.

Species[a]	Size (Mb)	# of ORFs	# of Exons	Chromo-some	GC content & Skews	Inter-Pro	ESTs	Source[b]
Fungi (Kingdom)[f]								
Ascomycota (Phylum)								
Pezizomycotina (Subphylum)								
A: *Botrytis cinerea* T: *Botryotinia fuckeliana*	42.7	16,448	43,358	N	Y	Y	N	BI
Sclerotinia sclerotiorum	38.3	14,522	40,623	N	Y	Y	N	BI
Aspergillus clavatus	27.9	9,121	27,959	N	Y	Y	N	BI
Aspergillus flavus	36.8	12,587	40,971	N	Y	Y	N	BI
Aspergillus fumigatus AF293	29.4	9,887	28,164	8	Y	Y	N	TIGR
Aspergillus fumigatus A1163	29.2	9,929	29,094	N	Y	Y	N	TIGR
A: *Aspergillus nidulans* T: *Emericella nidulans*	30.1	10,701	35,525	8	Y	Y	N	BI
Aspergillus niger ATCC1015	37.2	11,200	34,971	N	Y	Y	N	JGI
Aspergillus niger CBS513.88	34.0	14,086	50,371	8	Y	Y	N	NCBI

A: *Aspergillus oryzae* T: *Eurotium oryzae*	37.1	12,063	35,319	N	Y	Y	N	DOGAN
Aspergillus terreus	29.3	10,406	33,116	N	Y	Y	N	BI
A: *Aspergillus fischerianus* T: *Neosartorya fischeri*[d]	32.6	10,403	N	N	Y	N	N	BI
Penicillium marneffei	28.5	N	N	N	Y	N	N	TIGR
Coccidioides immitis RS	28.9	10,457	36,137	N	Y	Y	N	BI
Coccidioides immitis H538.4	27.7	10,663	34,503	N	Y	Y	N	BI
Coccidioides immitis RMSCC 2394	28.9	10,408	34,807	N	Y	Y	N	BI
Coccidioides immitis RMSCC 3703	27.7	10,465	33,931	N	Y	Y	N	BI
Coccidioides posadasii Silveria	27.5	10,125	33,520	N	Y	Y	N	BI
Coccidioides posadasii C735	27.0	N	N	N	Y	N	N	BI
Coccidioides posadasii CPA0001	28.7	N	N	N	Y	N	N	BI
Coccidioides posadasii CPA0020	27.3	N	N	N	Y	N	N	BI
Coccidioides posadasii CPA0066	27.7	N	N	N	Y	N	N	BI
Coccidioides posadasii RMSCC 1037	26.7	N	N	N	Y	N	N	BI
Coccidioides posadasii RMSCC 1038	26.2	N	N	N	Y	N	N	BI

Coccidioides posadasii RMSCC 1040	26.5	N	N	N	Y	N	N	BI
Coccidioides posadasii RMSCC 2133	27.9	N	N	N	Y	N	N	BI
Coccidioides posadasii RMSCC 3488	28.1	9,964	33,484	N	Y	Y	N	BI
Coccidioides posadasii RMSCC 3700	25.5	N	N	N	Y	N	N	BI
Paracoccidioides brasiliensis Pb01	33.0	9,136	37,310	N	Y	Y	N	BI
Paracoccidioides brasiliensis Pb03	29.1	9,264	31,468	N	Y	Y	N	BI
Paracoccidioides brasiliensis Pb18	30.0	8,741	33,239	N	Y	Y	N	BI
Blastomyces dermatitidis	61.8	N	N	N	Y	N	N	WGSC
A: *Histoplasma capsulatum* G217B T: *Ajellomyces capsulatus* G217B	41.3	8,038	26,711[h]	N	Y	Y	N	WGSC
A: *Histoplasma capsulatum* G186AR T: *Ajellomyces capsulatus* G186AR	29.9	7,454	24,562[h]	N	Y	Y	N	WGSC
A: *Histoplasma capsulatum* NAm1 T: *Ajellomyces capsulatus* NAm1	33.0	9,349	32,844	N	Y	Y	N	BI
A: *Arthroderma gypseum* T: *Microsporum gypseum*	23.3	N	N	N	Y	N	N	BI
Ascosphaera apis	21.6	N	N	N	Y	N	N	BGM
Uncinocarpus reesii	22.3	7,798	24,094	N	Y	Y	N	BI
Chaetomium globosum[d]	34.9	11,124	N	N	Y	N	N	BI

Epichloe festucae	27.0	N	N	N	Y	N	N	OU
A: *Fusarium graminearum* PH-1 T: *Gibberella zeae* PH-1	36.6	13,321	37,549	N	Y	Y	N	BI
A: *Fusarium graminearum* GZ3639 T: *Gibberella zeae* GZ3639[c]	15.1	6,694	11,692[h]	N	Y	Y	N	BI
Fusarium oxysporum f. sp. lycopersici 4286	61.4	17,608	47,051	15	Y	Y	N	BI
A: *Fusarium verticillioides* 7600 T: *Gibberella moniliformis* 7600	41.9	14,155	39,058	N	Y	Y	N	BI
A: *Fusarium solani* MPVI T: *Nectria haematococca* MPVI	51.3	15,707	48,387	N	Y	Y	N	JGI
A: *Pyricularia oryzae* 70-15 T: *Magnaporthe oryzae* 70-15	41.6	12,841	34,189	7	Y	Y	Y	BI
A: *Pyricularia oryzae* 70-15 chromosome 7 T: *Magnaporthe oryzae* 70-15 chromosome 7	4.0	1,122	3,289	1	Y	Y	N	
Neurospora crassa OR74A	39.2	9,842	27,188	8	Y	Y	N	BI
Podospora anserina DSM980	35.7	10,596	24,437	9	Y	Y	N	IGM
Trichoderma atroviride IMI206040	36.0	11,100	32,563	N	Y	Y	N	JGI
A: *Trichoderma reesei* QM6a T: *Hypocrea jecorina* QM6a	33.5	9,129	27,891	N	Y	Y	N	JGI
A: *Trichoderma virens* Gv29-8 T: *Hypocrea virens* Gv29-8	38.8	11,643	34,673	N	Y	Y	N	JGI
Talaromyces stipitatus ATCC 10500	35.5	N	N	N	Y	N	N	TIGR
Verticillium dahliae VaLs. 17	33.9	10,575	29,736	N	Y	N	N	BI

Verticillium albo-atrum VaMs. 102	32.9	10,239	28,842	N	Y	N	N	BI
Alternaria brassicicola	32.0	N	N	N	Y	N	N	WGSC
A: *Bipolaris maydis* T: *Cochliobolus heterostrophus* C5	34.9	9,633	28,007	N	Y	N	N	JGI
Pyrenophora tritici-repentis	38.0	12,169	32,717	N	Y	Y	N	BI
A: *Septoria tritici* T: *Mycosphaerella graminicola*	41.9	11,395	30,629	N	Y	Y	N	JGI
A: *Paracercospora fijiensis* T: *Mycosphaerella fijiensis*	73.4	10,313	25,289	N	Y	Y	N	JGI
A: *Stagonospora nodorum* T: *Phaeosphaeria nodorum*	37.2	16,597	44,017	N	Y	Y	N	BI
Saccharomycotina (Subphylum)								
Candida albicans SC5314	14.3	6,090	6,624	N	Y	Y	N	SGTC
Candida albicans WO-1	14.4	6,160	6,395	N	Y	Y	N	BI
Candida dubliniensis[d]	14.5	6,027	N	N	Y	N	N	SI
Candida glabrata CBS138	12.3	5,165	5,249	N	Y	Y	N	CBS
A: *Candida guilliermondii* T: *Pichia guilliermondii*	10.6	5,920	5,935	N	Y	Y	N	BI
Candida lusitaniae	12.1	5,941	5,956	N	Y	Y	N	BI
Candida parapsilosis	13.1	5,733	5,733	N	Y	Y	N	BI

Candida tropicalis	14.7	6,258	6,292	N	Y	Y	N	BI
Candida tropicalis[g]	2.1	N	N	N	Y	N	N	GS
Ashbya gossypii	8.8	4,717	4,943	7	Y	Y	N	NCBI
Debaryomyces hansenii	12.2	6,354	6,710	7	Y	Y	N	CBS
Debaryomyces hansenii[g]	2.3	N	N	N	Y	N	N	GS
A: *Candida sphaerica* T: *Kluyveromyces lactis*	10.7	5,327	5,461	N	Y	Y	N	GS
A: *Candida sphaerica* T: *Kluyveromyces lactis*[g]	5.1	N	N	N	Y	N	N	GS
A: *Candida kefyr* T: *Kluyveromyces marxianus*[g]	2.0	N	N	N	Y	N	N	GS
Kluyveromyces polysporus DSM70294	14.7	5,367	5,524	N	Y	Y	N	SIG
Kluyveromyces thermotolerans[g]	2.2	N	N	N	Y	N	N	GS
Kluyveromyces waltii	10.9	4,935	5,395	N	Y	Y	N	BI
Lodderomyces elongisporus	15.5	5,796	5,856	N	Y	Y	N	BI
Saccharomyces bayanus MCYC 623	11.5	9,385	9,385	N	Y	Y	N	BI
Saccharomyces bayanus 623-6C YM4911	11.9	4,966	4,966	N	Y	Y	N	WGSC
Saccharomyces bayanus var. uvarum[g]	4.5	N	N	N	Y	N	N	GS

Saccharomyces castellii	11.4	4,677	4,677	N	Y	Y	N	WGSC
A: *Candida robusta* S288C T : *Saccharomyces cerevisiae* S288C	12.2	6,692	7,042	16	Y	Y	N	SGD
A: *Candida robusta* RM11-1a T : *Saccharomyces cerevisiae* RM11-1a	11.7	5,383	5,988	N	Y	Y	N	BI
A: *Candida robusta* YJM789 T : *Saccharomyces cerevisiae* YJM789	12.0	5,471	6,153	N	Y	Y	N	SI
Saccharomyces exiguus[g]	2.0	N	N	N	Y	N	N	GS
Saccharomyces kluyveri	11.0	2,968	2,968	N	Y	Y	N	WGSC
Saccharomyces kluyveri[g]	2.2	N	N	N	Y	N	N	GS
Saccharomyces kudriavzevii	11.2	3,768	3,768	N	Y	Y	N	WGSC
Saccharomyces mikatae	11.5	9,016	9,016	N	Y	Y	N	BI
Saccharomyces mikatae	10.8	3,100	3,100	N	Y	Y	N	WGSC
Saccharomyces paradoxus	11.9	8,939	8,939	N	Y	Y	N	BI
Saccharomyces servazzii[g]	2.0	N	N	N	Y	N	N	GS
Pichia angusta[g]	4.5	N	N	N	Y	N	N	GS
Pichia stipitis	15.4	5,839	8,428	N	Y	Y	N	JGI
Pichia sorbitophila[g]	3.8	N	N	N	Y	N	N	GS

A: *Candida lipolytica* T : *Yarrowia lipolytica*	20.5	6,524	7,264	6	Y	Y	N	CBS
A: *Candida lipolytica* T: *Yarrowia lipolytica*[g]	4.6	N	N	N	Y	N	N	GS
Zygosaccharomyces rouxii[g]	4.1	N	N	N	Y	N	N	GS
Taphrinomycotina (Subphylum)								
Pneumocystis carinii[c,d]	6.3	4,020	N	N	Y	N	N	SI
Schizosaccharomyces japonicus	11.3	5,172	10,321	N	Y	Y	N	BI
Schizosaccharomyces pombe	12.6	5,005	9,869	3	Y	Y	N	GeneDB
Schizosaccharomyces octosporus	11.2	4,925	10,168	N	Y	N	N	BI
Basidiomycota (Phylum)								
Agricomycotina (Subphylum)								
Postia placenta	90.9	17,173	116,596	N	Y	Y	N	JGI
T: *Phanerochaete chrysosporium* A: *Sporotrichum pruinosum*	30.0	10,048	58,746	N	Y	Y	N	JGI
Coprinus cinereus	36.3	13,544	72,887	N	Y	Y	N	BI
Laccaria bicolor	64.9	20,614	111,290	N	Y	Y	N	JGI
A: *Cryptococcus neoformans* Serotype A T: *Filobasidiella neoformans* Serotype A	19.5	7,302	43,325	20	Y	Y	N	BI

A: *Cryptococcus neoformans* Serotype B T: *Filobasidiella neoformans* Serotype B	19.0	6,870	40,589	N	Y	Y	N	NCBI
A: *Cryptococcus neoformans* Serotype D B-3501A T: *Filobasidiella neoformans* Serotype D B-3501A	18.5	6,431	40,942	N	Y	Y	N	SGTC
A: *Cryptococcus neoformans* Serotype D JEC21 T: *Filobasidiella neoformans* Serotype D JEC21	19.1	6,475	40,811	N	Y	Y	N	SGTC
Pucciniomycotina (Subphylum)								
Sporobolomyces roseus	21.2	5,536	39,911	N	Y	Y	N	JGI
Puccinia graminis	88.7	20,567	95,838	N	Y	Y	N	BI
Ustilaginomycotina (Subphylum)								
Malassezia globosa CBS7966	9.0	4,286	4,286	N	Y	N	N	PGC
Malassezia restricta CBS7877	4.6	N	N	N	Y	N	N	PGC
Ustilago maydis 521	19.7	6,689	11,589	N	Y	Y	N	BI
Ustilago maydis FB1	19.7	6,950	10,310[h]	N	Y	Y	N	BI
Chytridiomycota (Phylum)								
Batrachochytrium dendrobatidis JEL423	23.9	8,818	38,551	N	Y	Y	N	BI
Batrachochytrium dendrobatidis JAM81	24.3	8,732	37,423	N	Y	Y	N	JGI
Mucoromycotina (Subphylum *incertae sedis*)								

Rhizopus oryzae	46.1	17,467	57,981	N	Y	Y	N	BI
Phycomyces blakesleeanus	55.9	14,792	71,502	N	Y	Y	N	JGI
Microsporidia (Phylum)								
Encephalitozoon cuniculi	2.5	1,996	2,002	N	Y	Y	N	GS
Antonospora locustae[d]	6.1	2,606	N	N	Y	N	N	JBPC
Stramenopila (Kingdom)[e]								
Peronosporomycota (Phylum)								
Phytophthora infestans[d]	228.5	22,658	N	N	Y	N	N	BI
Phytophthora ramorum	66.7	16,066	40,639	N	Y	Y	N	JGI
Phytophthora sojae	86.0	19,276	53,552	N	Y	Y	N	JGI
Hyaloperonospora parasitica	83.8	14,789	24,907	N	Y	Y	N	VBI
Subtotal in fungi and oomycetes	3,536.9	924,343	2,605,087	14	129	88	1	
Chloroplastida (Kingdom)[e]								
Streptophyta (Phylum)								
Arabidopsis thaliana	119.2	28,581	150,369	5	Y	Y	N	TAIR

Carica papaya	271.7	N	N	N	Y	N	N	PGSC
Lycopersicon esculentum[c]	39.9	8,725	29,707	N	Y	Y	N	SOL
Medicago truncatula	278.7	38,334	122,889	8	Y	Y	N	MTGSP
Oryza sativa var. Indica[d]	426.3	49,710	N	N	N	N	N	BGI
Oryza sativa var. Japonica	372.0	66,710	319,140	12	N	Y	N	IRGSP
Populus trichocarpa	427.3	45,555	193,687	N	N	Y	N	JGI
Ricinus communis[d]	362.4	38,613	N	N	N	N	N	TIGR
Selaginella moellendorffii	212.8	22,285	124,645	N	N	Y	N	JGI
Sorghum bicolor	738.5	36,338	165,149	11	N	Y	N	JGI
Vitis vinifera	497.5	30,434	149,351	19	Y	Y	N	GS
Zea mays[d]	2,314.7	420,732	N	N	N	N	N	MGSP
Metazoa (Kingdom)								
Arthropoda (Phylum)								
Apis mellifera	235.2	11,062	71,496	N	Y	N	N	HBGP
Acyrthosiphon pisum	446.6	N	N	N	N	N	N	BCM

Bombyx mori	397.7	21,302	82,381	N	N	N	N	BGI
Drosophila ananassae	230.9	15,276	56,595	N	Y	N	N	Flybase
Drosophila erecta	152.7	15,324	56,924	N	Y	N	N	Flybase
Drosophila grimshawi	200.5	15,270	56,647	N	Y	N	N	Flybase
Drosophila melanogaster	168.5	20,923	96,745	N	Y	N	N	Flybase
Drosophila mojavensis	193.8	14,849	55,013	N	Y	N	N	Flybase
Drosophila persimilis	188.4	17,235	59,116	N	Y	N	N	Flybase
Drosophila pseudoobscura	152.7	9,868	39,902	N	Y	N	N	Flybase
Drosophila sechellia	166.6	16,884	58,584	N	Y	N	N	Flybase
Drosophila simulans	137.8	15,983	54,294	N	Y	N	N	Flybase
Drosophila virilise	206.0	14,680	55,005	N	Y	N	N	Flybase
Drosophila willistoni	235.5	15,816	56,641	N	Y	N	N	Flybase
Drosophila yakuba	165.7	15,423	59,098	N	Y	N	N	Flybase
Glossina morsitans	205.7	N	N	N	N	N	N	TIGR
Nasonia vitripennis	239.6	27,957	98,570[h]	N	N	N	N	BCM

Tribolium castaneum	152.1	14,274	58,381[h]	N	Y	N	N	BCM
Nematoda (Phylum)								
Caenorhabditis elegans	100.3	26,902	175,232	7	Y	N	N	Wormbase
Caenorhabditis briggsae[d]	108.5	20,669	N	N	Y	N	N	Wormbase
Caenorhabditis remanei	145.4	N	N	N	Y	N	N	Wormbase
Vertebrata (Phylum)								
Homo sapiens Celera assembly	2,828.4	28,057	273,999	N	Y	N	N	NCBI
Homo sapiens HuRef assembly	2,843.9	27,937	273,135	N	Y	N	N	NCBI
Homo sapiens NCBI Reference	2,870.8	29,319	284,553	N	Y	N	N	NCBI
Homo sapiens	3,418.7	33,869	452,099	29	Y	N	N	Ensembl
Subtotal for non-fungal genomes	22,253.0	1,214,896	3,729,347	7	26	8	0	
Total	25,789.9	2,39,239	6,334,434	21	155	96	1	

[a]A indicates anamorph name and T presents teleomorph name of fungi.

[b]SGTC, Stanford Genome Technology Center; SI, Sanger Institute; CBS, Center For Biological Sequences; BI, Broad Institute; WGSC, Washington University Genome Sequencing Center; JGI, DOE Joint Genomic Institute; DOGAN, Database Of the

Genomes Analyzed at Nite; GS, Genoscope; IGM, Instituté de Génétique et Microbiologie; TAIR, The Arabidopsis Information Resource; IRGSP, International Rice Genome Sequencing Project; BGI, Beijing Genome Institute; VGI, Virginia Bioinformatics Institute; SIG, Trinity College Dublin, Smurfit Institute of Genetics; JBPC, Josephine Bay Paul Center for Comparative Molecular Biology and Evolution; MTGSP, *Medicago Truncatula* Genome Sequencing Project; HBGP, Honey Bee Genome Project; BCM, Baylor College of Medicine; MGSP, Maize Genome Sequencing Project; PGSC, Papaya Genome Sequencing Consortium; BGM, Baylor College of Medicine; OU, Oklahoma University; PGC, Procter & Gamble Co.

[c]Incomplete coverage of genome information

[d]Insufficient exon/intron information

[e]Taxonomy based on (Adl *et al.*, 2005)

[f]Taxonomy based on (Hibbett *et al.*, 2007)

[g]Sequences from Random Sequence Tag (RST)

[h]ORFs and exons were predicted by AUGUSTUS 2.0.3 with species-specific training datasets (Stanke *et al.*, 2004).

‘Y’ indicates the existence of information in each field, and ‘N’ indicates the lack of information.

comparison of genomic contexts of any orthologous genes among species. In addition, SNUGB displays the InterPro terms assigned to each ORF as well as the genomic regions where expressed sequence tags (ESTs) are matched. With these functionalities, SNUGB will serve as a powerful platform supporting comprehensive fungal comparative genomics.

RESULTS

I. Data processing via an automated pipeline and the function of Positional Database

Positional information of functional/structural units, that are present on individual contigs/chromosomes, such as the start and stop sites of ORFs and exons/introns, was collected from the data warehouse of CFGP and stored in the Position Database of SNUGB. New types of data, such as Simple Sequence Repeats (SSRs) on the genome, can be easily added to the Positional Database for visualization via SNUGB. Along with the positional information, for each data, data type (e.g., ORFs), primary key, and any additional information were saved into the partitioned tables, which were designed for enhancing the speed of data retrieval. Through the primary key, SNUGB can display detailed information of each data (e.g., sequences) stored at external sources. Considering the large number of available fungal genome sequences and those that are currently being sequenced, in addition to this data standardization scheme, a standardized pipeline for data extraction and management is needed to organize the data and to ensure orderly expansion of SNUGB.

The pipeline developed for SNUGB processes each genome data set via the following steps. Firstly, once whole genome sequences are deposited in the data warehouse of CFGP, the integrity of genome information, such as the position information of functional/structural units, is inspected. Several properties of the whole genome, such

as the length and the GC content, are calculated. Secondly, the GC content, AT-skew, and CG-skew are calculated via 50-bp sliding windows with 20bp steps. Thirdly, for each gene, three types of sequence information, including coding sequences (sequences from the start to stop codon without introns), gene sequences (sequences from the start to stop codon with introns), and transcript sequences (sequences from the transcription start site to end site without intron sequence), if transcript information is available, are generated based on the genome annotation information. Fourthly, all data generated in the previous steps are transferred into the Position Database to support graphical representation of these features. Fifthly, if the genome has chromosomal map information, including genetic map and optical map, this information is converted into a standardized format and stored in SNUGB for graphical representation via Chromosome Viewer. Lastly, after subjecting all ORFs in the genome through the InterPro Scan (Mulder *et al.*, 2007), the genomic position of each domain predicted by the InterPro Scan is calculated and stored into the Position Database.

II. Modular design of SNUGB facilitates its application

To facilitate the efficient implementation of SNUGB in diverse genomics platforms, a modular design was used for its application programming interface (API). Through API, a diagram showing genome features in a selected region can be created using only their chromosomal positions and display options. Two recent publications illustrate the utility of this design: T-DNA Analysis Platform (TAP; http://tdna.snu.ac.kr/) provides the GC content and AT skew around T-DNA insertion sites on the chromosomes of *Magnaporthe oryzae* via a mini genome browser (MiniGB) supported by SNUGB (Choi *et al.*, 2007). The chromosomal distribution pattern of T-DNA insertion sites in *M. oryzae* (http://atmt.snu.ac.kr/) was also displayed using SNUGB (Jeon *et al.*, 2007). Fungal Cytochrome P450 Database (FCPD; http://p450.riceblast.snu.ac.kr/) (Park *et al.*, 2008a) employs SNUGB to present the chromosomal distribution pattern and contexts of cytochrome P450 genes in fungal genomes. Two databases, FED (http://fed.snu.ac.kr/) and FSD (http://fsd.snu.ac.kr/), utilize SNUGB for presenting the genomic context of the region matched to EST and secreted proteins, respectively. Moreover, Insect Mitochondrial Genome Database (IMGD; http://www.imgd.org/) (Lee *et al.*, 2009) and Systemic Protein Identification of Mutated Protein (SysPIMP; http://syspimp.starflr.info/) (Xi *et al.*, 2009) also adopted SNUGB for data presentation. These examples illustrate the utility of SNUGB.

III. Properties of the fungal/oomycete genomes archived in SNUGB

Among the 91 fungal/oomyvete species (129 genome datasets) covered by SNUGB (Table 1), 70 species (104 genome datasets; 81%) belong to the phylum Ascomycota, and 10 species (14 genome datasets; 11%) belong to the phylum Basidiomycota. In contrast, the phyla Chytridiomycota and Micosporidia are represented only by one (2 datasets) and two species (both belong to the subphylum Mucoromycotina), respectively. Four oomycete genomes, derived from *Phytophthora* and *Hyaloperonospora* species, are available for comparison with fungal genomes. Although oomycetes belong to the kingdom Stramenophla and show closer phylogenetic relationships to algae and diatoms than fungi (Forster *et al.*, 1990), due to their morphological similarities to fungi, they have been traditionally grouped with fungi.

The datasets that cover the whole genome (113 out of the 129 datasets) were analyzed to investigate genome properties. The average size of the genomes is 30.48 Mb which is one-tenth of plant genomes (317.07 Mb in the phylum Streptophyta) and one-seventh of insect genomes (215.36 Mb in the phylum Arthropoda) (Figure 1A). The fungal/oomycete genome sizes ranged from 2.5Mb (*Encephalitozoon cuniculi*) to 228.5Mb (*Phytophthora infestans*); the genome of *E. cuniculi* is shorter than that of *Escherichia coli* (4.6 Mb) (Blattner *et al.*, 1997), while the genome of *P. infestans* is much larger than the genomes of *Arabidopsis thaliana* (119.2Mb) (AGI, 2000) and *Caenorhabditis elegans* (100.5Mb) (CSC, 1998), indicating no clear relationship

between the genome size and the organismal complexity (Gregory, 2004). With regard to the average genome sizes in different taxon groups, the phylum Microsporidia, known as ancestral fungi, shows the smallest average size (4.28 Mb), while oomycetes show the largest at 116.20 Mb (Figure 1A). In the phylum Basidiomycota, which is large and very diverse, the degree of difference in average genome sizes within each of the represented subphyla is highest in the fungal kingdom: the ratios of standard deviation to the average length in three subphyla Agricomycotina, Pucciniomycotina, and Ustilaginomycotina are 71.95%, 86.93%, and 57.46%, respectively (Figure 1B). The subphylum Pucciniomycotina displays the largest size with large variation (Figure 1A and 1B), while two subphyla Saccharomycotina and Taphrinomycotina belonging to the phylum Ascomycota exhibit the relatively low degree of variations (Figure 1B), probably because only closely related species have been sequenced. Although the average genome sizes varied from group to group, ANOVA and TukeyHSD tests ($P < 0.05$) showed only the difference between fungi and oomycetes was significant (Figure 1A). The GC content of fungal genomes ranges from 32.523% (*Pneumocystis carinii* in subphylum Taphrinomycotina) to 56.968% (*Phanerochaete chrysosporium* in the subphylum Agricomycotina), while the GC content of plant and insect genomes ranges from 29.638% to 46.850% (Figure 1C). Although the coding regions exhibit higher GC contents than the rest of the genome, there is no relationship between the proportion of ORFs on the genome and the GC content of the whole genomes (linear regression; $R^2 = 0.04$; Figure 1C and 1D).

The number of total proteins encoded by each organism was once considered to reflect organism's characteristics (Thomas, 1971). Based on the size of total proteomes, all sequenced fungal and oomycete species were divided into three groups: The medium group contains the subphylum Pezizomycotina in Ascomycota and the subphyla Agricomycota and Puccinomycotina in Basidiomycota, the small group includes three subphyla Saccharomycotina, Taphrinomycotina, and Ustilagomycotina and the phylum Microsporidia, and the large group has the subphylum Mucoromycotina and the phylum Oomycota (ANOVA and TukeyHSD; $P < 0.05$; Figure 1E). This grouping shows that the number of total ORFs does not correlate with taxonomic positions at the phylum level, however, at the subphylum level, the correlation was high. For example, subphyla Sacchromycotina and Taphrinomycotina can be distinguishable from Pezizomycotina based on this character. The ORF density classified the sequenced species into three distinct groups, Oomycetes, Microsporidia and the rest, through ANOVA and TukeyHSD test ($P < 0.05$; Figure 1F). Taken together, these three indicators can be used to divide fungal subphyla/phyla. For example, the subphylum Pezizomycotina shows the medium-level of ORF number and ORF density, while the subphylum Sacchromycotina displays the low-level of ORF number but its ORF density is comparable to that of the subphylum Pezizomycotia. Both the number of ORFs and the ORF density are high for oomycetes, exhibiting a pattern different from fungi.

The number of exons per ORF was investigated, resulting in four groups (ANOVA and TukeyHSD test; $P<0.05$; Figure 1G). With the exception of the subphylum

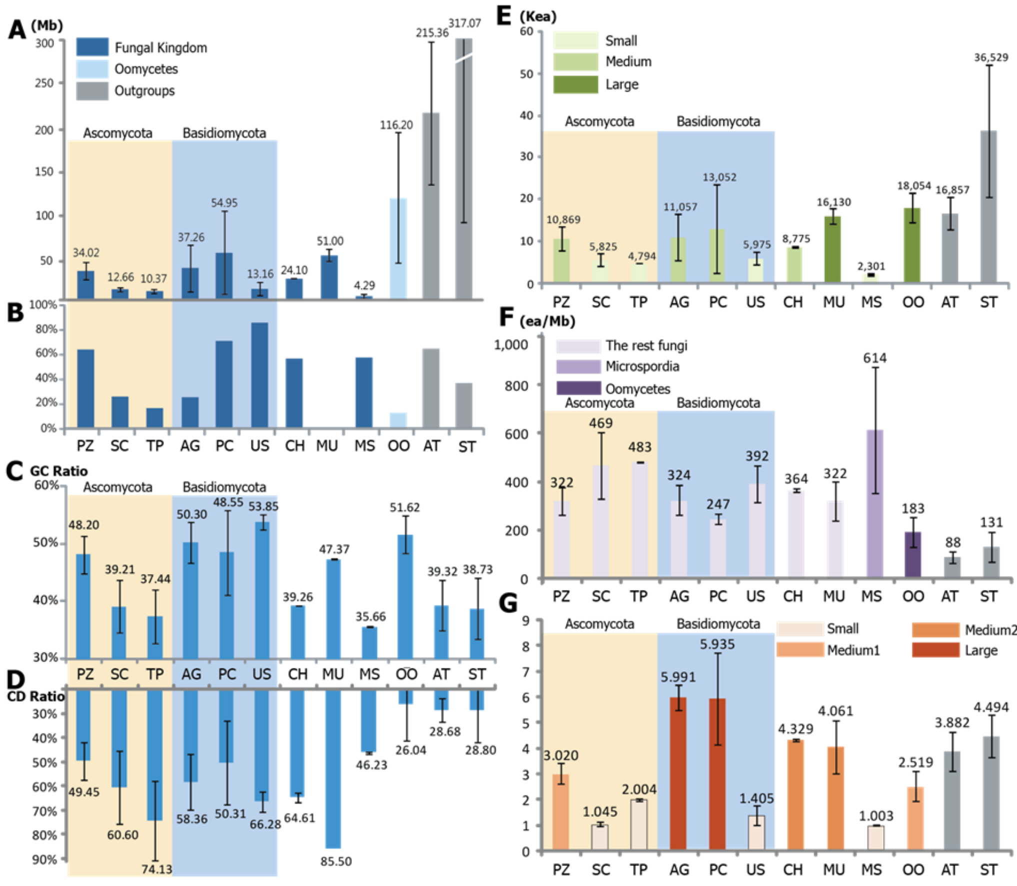


Figure 1. Characteristics of the 129 fungal and oomycetes genomes archived in SNUGB

In all graphs, the first six groups correspond to subphyla and the rests indicate phyla. Error bars indicate variation of data within each taxonomic group. Abbreviations for subphyla/phyla indicates that PZ; the subphylum Pezizomycotina, SC; the subphylum Saccharomycotina, TP; the subphylum Taphrinophycotina; AG; the subphylum Agricomycotina, PC; the subphylum Pucciniomycotina, US; the subphylum Ustilagomycotina, CH; the phylum Chytridiomycota, MU; the subphylum

Mucoromycotina, MS; the phylum Microsporidia, OO; oomycete (the phylum Peronosporomycota), AT; the phylum Arthropoda, and ST; the phylum Streptophyta. The last two phyla were used as outgroup. In graphs A, E, F, and G, each color of bar indicates distinct group supported by TurkeyHSD test. (A) Average genome size. (B) the ratio of variation of genome size to the average genome size. (C) Average GC ratio of each subphylum/phylum. (D) The percentage of coding regions to the genome length. (E) Average number of total ORFs. (F) The total number of ORFs per Mb (=ORF density). (G) The average exon number of each ORFs.

Ustilagomycotina, the phylum Basidiomycota exhibits the highest number (~6). The subphyla Sacchromycotina and Mycoromycotina show the lowest value (nearly 1), indicating that almost all their genes do not have introns.

IV. Comparison of genome sequences of multiple isolates within species

For 14 fungal species, two or more strains have been sequenced (Table 2). For some species, such as *Fusarium graminearum*, additional isolate(s) were sequenced only at a low coverage (e.g., 0.4x coverage for the second strain of *F. graminearum*); however, even such low-coverage provided some insights into the evolution of pathogenicity in this important cereal pathogen (Cuomo *et al.*, 2007). Except *Aspergillus niger*, *Histoplasma capsulatum, and Paracoccidioides brasiliensis*, all strains within same species showed less than 1 Mb variation in genome sizes (Table 2). Three isolates of *H. capsulatum* and *P. brasiliensis* showed approximately 1% difference in the GC content, whereas the degree of GC content variation among 11 strains of *Coccidioides posadasii* was only 0.5%. Four *Cryptococcus neoformans* strains, representing three different serotypes (A, B and D), showed around 0.3% variation in the GC content, and within a serotype (two serotype D strains) the difference was only 0.043% (Loftus *et al.*, 2005). Isolates of *Candida albicans*, *Sacchromyces bayanus,* and *Batrachochytrium dendrobatidis* showed only 0.01% variation in the GC content. These intraspecific variations of genome properties can be compared in detail via SNUGB.

Table 2. Basic properties of different strains of fungal genomes deposited in SNUGB.

Species	# of Strains	Genome size (Mb)	GC content (%)
Fungi (Kingdom)[e]			
Ascomycota (Phylum)			
Pezizomycotina (Subphylum)			
Aspergillus fumigatus	2	29.3 ± 0.1	49.672 ± 0.178
Aspergillus niger	2	35.6 ± 2.3	50.365 ± 0.012
Coccidioides immitis	4	28.3 ± 0.7	46.529 ± 0.514
Coccidioides posadasii	11	27.2 ± 0.9	46.839 ± 0.537
Histoplasma capsulatum	3	34.7 ± 5.9	44.430 ± 1.707
Paracoccidioides brasiliensis	3	30.7 ± 2.0	43.868 ± 0.930
Fusarium graminearum[a]	2	36.6	48.283
Saccharomycotina (Subphylum)			
Candida albicans	2	14.4 ± 0.1	33.462 ± 0.010
Saccharomyces cerevisiae	3	11.9 ± 0.3	38.252 ± 0.090
Saccharomyces bayanus	2	11.7 ± 0.3	40.196 ± 0.011
Saccharomyces mikatae[b]	2	11.1 ± 0.5	37.920 ± 0.315
Basidiomycota (Phylum)			
Agricomycotina (Subphylum)			
Cryptococcus neoformans	4	19.2 ± 0.2	48.251 ± 0.316
Ustilaginomycotina (Subphylum)			
Ustilago maydis	2	19.7 ± 0.0	53.995 ± 0.045
Chytridiomycota (Phylum)			
Batrachochytrium dendrobatidis	2	24.1 ± 0.3	39.261 ± 0.011

Chloroplastida (Kingdom)			
Charophyta (Phylum)			
Oryza sativa	2	399.2 ± 38.4	43.530 ± 0.046
Vertebrata (Phylum)			
Vertebrata (Phylum)			
Homo sapiens	4	3,052.2 ± 409.3	40.878 ± 0.042

[a]One of strains are incomplete whole genome sequences, so that standard deviation of genome length and GC content are not calculated.

[b]Same strain but different version of assembly

V. Update of SNUGB

The number of on-going fungal genome sequencing projects is approximately 40 (http://fungalgenomes.org/wiki/Fungal_Genome_Links). 37 strains of *S. cerevisiae* and 25 strains of *S. paradoxus* were already sequenced and released by the Sanger institute (http://www.sanger.ac.uk/Teams/Team71/durbin/sgrp/index.shtml), indicating that more than 100 additional fungal genomes will be available soon. Next generation high throughput sequencing technologies, such as GS Flx, Solexa, and SOLiD (Ellegren, 2008; Shiu and Borevitz, 2008), will further accelerate the rate of fungal genome sequencing, emphasizing the importance of frequently updating SNUGB. With the aid of the developed pipeline, the SNUGB will be updated whenever new fungal genome sequences have been publicly released with annotation information. A notice for updated genomes will be posted on the SNUGB web site.

VI. Functions and tools

Taxonomy browser

To support selection of species of interests based on their taxonomic positions, a web-based tool, named as the taxonomy browser, was developed. Considering an anticipated increase in comparing genome sequences and features across multiple species to investigate evolutionary questions at the genome scale, such a tool is necessary to provide an overview of the taxonomic positions of the sequenced species and their evolutionary relationships with other fungi to users of SNUGB and to assist them in selecting appropriate species for comparative analyses. The taxonomy browser provides two methods for accessing the data archived in SNUGB, one of which is text-search using species name (Figure 2A). The other method is using the taxonomical hierarchy (i.e., tree of life). When a user clicks a specific taxon (e.g., phylum), taxonomy browser will present all subgroups within the chosen taxon for further selection (Figure 2B).

Chromosome viewer and Contig/ORF browser

Three different methods can be used to access genomic information. For those with chromosomal map data (21 species), their chromosomal maps can be displayed via Chromosome viewer (Figure 3A). The following color scheme was used to denote the level of completeness: i) chromosome constructed using genetic or optical map data (with gaps) as blue (Chromosomes 1 to 7 of *M. oryzae*; Figure 3A), ii) chromosome map based on a combination of sequences and genetic/optical map data as pink (e.g.,

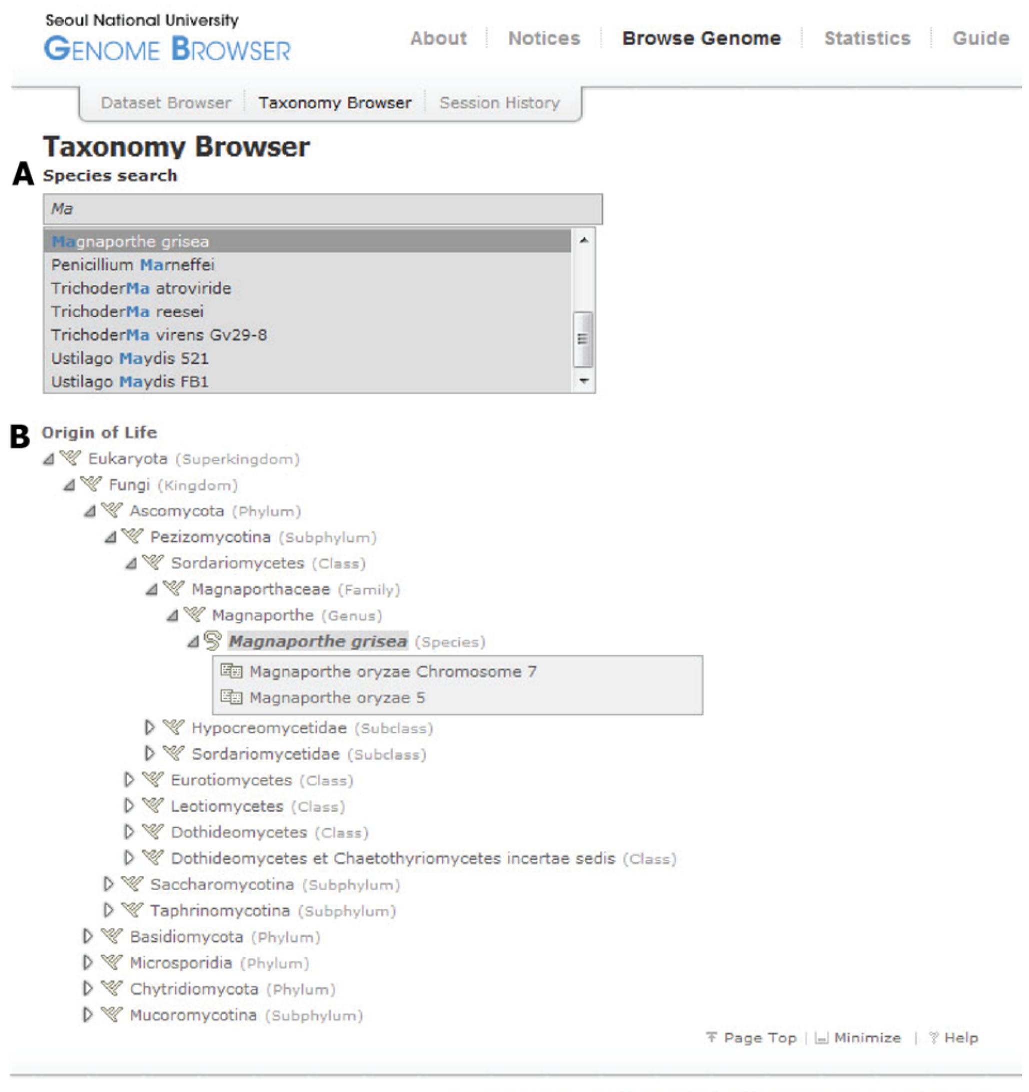


Figure 2. Taxonomy browser

A screenshot of data generated using Taxonomy browser is shown. (A) Search interface by species name shows a list of species along with inserted string. (B) Taxonomical tree shows a lineage of the chosen species and its genome datasets deposited in SNUGB.

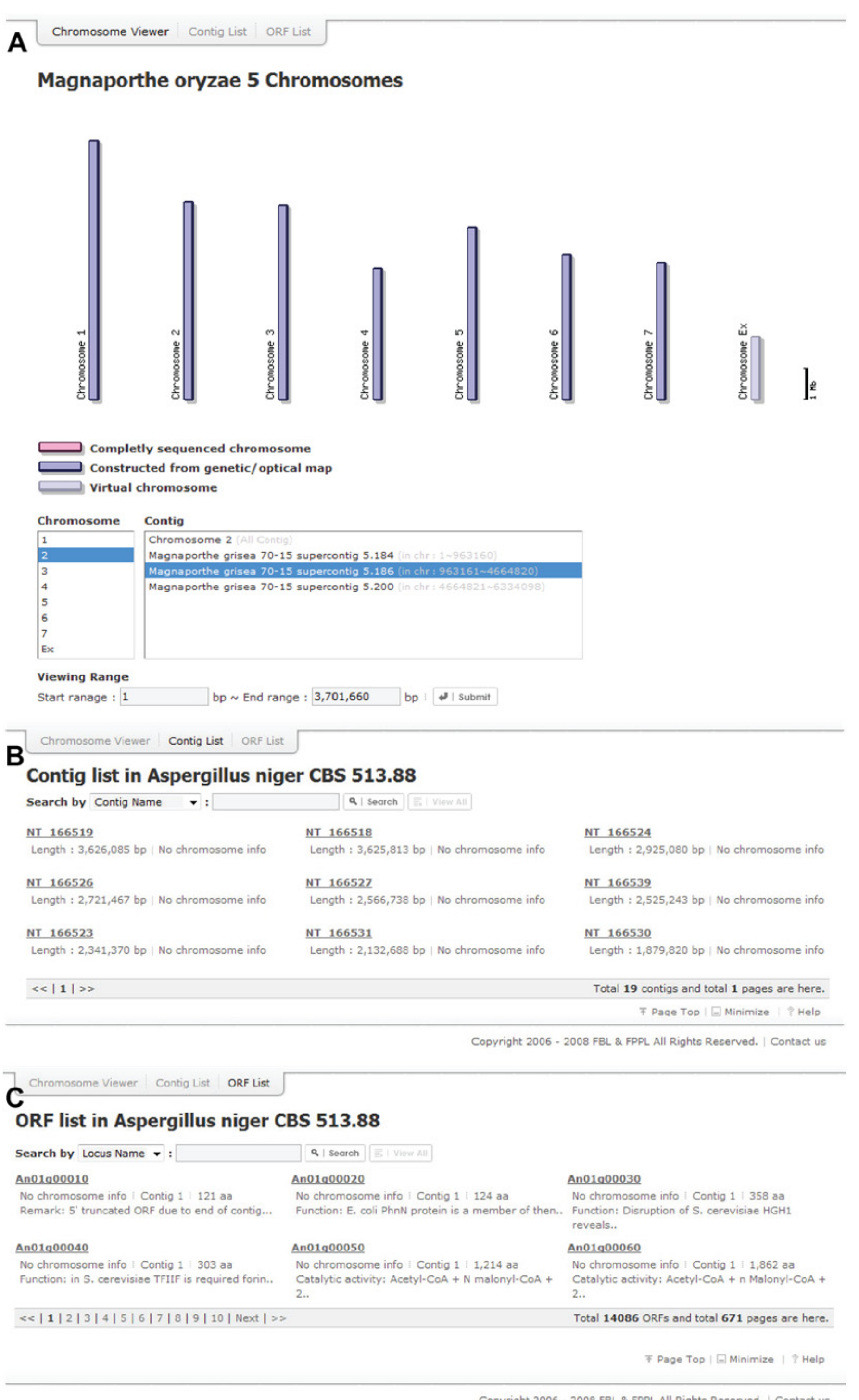

A
Chromosome Viewer
Contig List
ORF List
Magnaporthe oryzae 5 Chromosomes
Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome Ex
Completly sequenced chromosome
Constructed from genetic/optical map
Virtual chromosome
Chromosome
Contig
Chromosome 2
Magnaporthe grisea 70-15 supercontig 5.184
Magnaporthe grisea 70-15 supercontig 5.186
Magnaporthe grisea 70-15 supercontig 5.200
Viewing Range
Start ranage : 1 bp ~ End range : 3,701,660 bp
Submit
B
Contig list in Aspergillus niger CBS 513.88
Search by Contig Name
Search
NT_166519 Length : 3,626,085 bp | No chromosome info
NT_166518 Length : 3,625,813 bp | No chromosome info
NT_166524 Length : 2,925,080 bp | No chromosome info
NT_166526 Length : 2,721,467 bp | No chromosome info
NT_166527 Length : 2,566,738 bp | No chromosome info
NT_166539 Length : 2,525,243 bp | No chromosome info
NT_166523 Length : 2,341,370 bp | No chromosome info
NT_166531 Length : 2,132,688 bp | No chromosome info
NT_166530 Length : 1,879,820 bp | No chromosome info
Total 19 contigs and total 1 pages are here.
Page Top | Minimize | Help
Copyright 2006 - 2008 FBL & FPPL All Rights Reserved. | Contact us
C
ORF list in Aspergillus niger CBS 513.88
Search by Locus Name
An01g00010 No chromosome info | Contig 1 | 121 aa Remark: 5' truncated ORF due to end of contig...
An01g00020 No chromosome info | Contig 1 | 124 aa Function: E. coli PhnN protein is a member of then..
An01g00030 No chromosome info | Contig 1 | 358 aa Function: Disruption of S. cerevisiae HGH1 reveals..
An01g00040 No chromosome info | Contig 1 | 303 aa Function: in S. cerevisiae TFIIF is required forin..
An01g00050 No chromosome info | Contig 1 | 1,214 aa Catalytic activity: Acetyl-CoA + N malonyl-CoA + 2..
An01g00060 No chromosome info | Contig 1 | 1,862 aa Catalytic activity: Acetyl-CoA + n Malonyl-CoA + 2..
<< | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Next | >>
Total 14086 ORFs and total 671 pages are here.
Page Top | Minimize | Help
Copyright 2006 - 2008 FBL & FPPL All Rights Reserved. | Contact us

Figure 3. Chromosome viewer, Contig Viewer and ORF Viewer

(A) The chromosome viewer displays seven chromosomes of *M. oryzae* with a size indicator at the right side. At the bottom, the interface allows for jumping directly to a specific region by selecting chromosome/contigs and its position. (B) The contig viewer provides a list of contigs with its length. Through this interface, contigs can be searched by name. (C) The ORF viewer presents the names and lengths of ORFs with search function.

chromosomes of *A. niger*), and iii) unassigned contigs (labeled as Chromosome Ex of *M. oryzae*; Figure 3A) as light blue. For the species without chromosomal map information, SNUGB provides the contig and ORF browsers, which display the name of contig and ORFs, respectively, and allow users to search them using their names (Figures 3B and 3C).

Graphical Browser with six defferent display formats

Gene annotation information in a selected area of chromosome or contig, such as transcripts, ORFs and exon/intron structure, and InterPro domains (Mulder *et al.*, 2007), can be displayed through three formats: i) the 'single' format shows these features as bars; ii) the 'squish' format displays them via color-coded diagrams without description; and iii) the 'pack' format presents them as small color-coded icons with description (Figure 4A). These graphical formats were also used by UCSC Genome Browser (Kuhn *et al.*, 2007). In addition, the GC content and AT skew information for individual chromosomes can be displayed via three formats: i) color-coded bar graph, ii) line, and iii) dotted lines along with a description of data (Figure 4B). For species with EST data (Table 1), the genomic region corresponding to each EST sequence can be displayed along with ORF and InterPro domains to help users identify predicted gene structure and expressed regions (see Figure 4A). Presentation of these data is supported by Fungal Expression Database (http://fed.snu.ac.kr/).

A

Magnaporthe oryzae 5

Chromosome 1 (Total Length : 8,322,655bp)

Magnifying Option Zoom In +2x +4x Select Range Zoom Out -1.5x -3x View All

Range Option << < Start 1,566,263 bp End 1,786,654 bp > >> Refresh

1600000 1650000 1700000 1750000

CONTIG REGION INFORMATION

bar

GENE REGION INFORMATION

squash

GENE REGION INFORMATION

GENE REGION INFORMATION

pack

EST REGION

INTERPRO TERM

Viewing Option Contig Region Squish ORF Pack View Exon/Intron Option Display Refresh

Additional Viewing Option Refresh

AT Skew None CG Skew None EST Region Pack

InterPro Term Pack GC Ratio None

B

Magnaporthe oryzae 5

Chromosome 1 (Total Length : 8,322,655bp)

Magnifying Option Zoom In +2x +4x Select Range Zoom Out -1.5x -3x View All

Range Option << < Start 1,566,263 bp End 1,786,654 bp > >> Refresh

1600000 1650000 1700000 1750000

graph

INFORMATION

AT Skew

line

GC Skew

S. line

GC Ratio

Viewing Option Contig Region Squish ORF None Refresh

Additional Viewing Option Refresh

AT Skew Graph CG Skew Line EST Region None

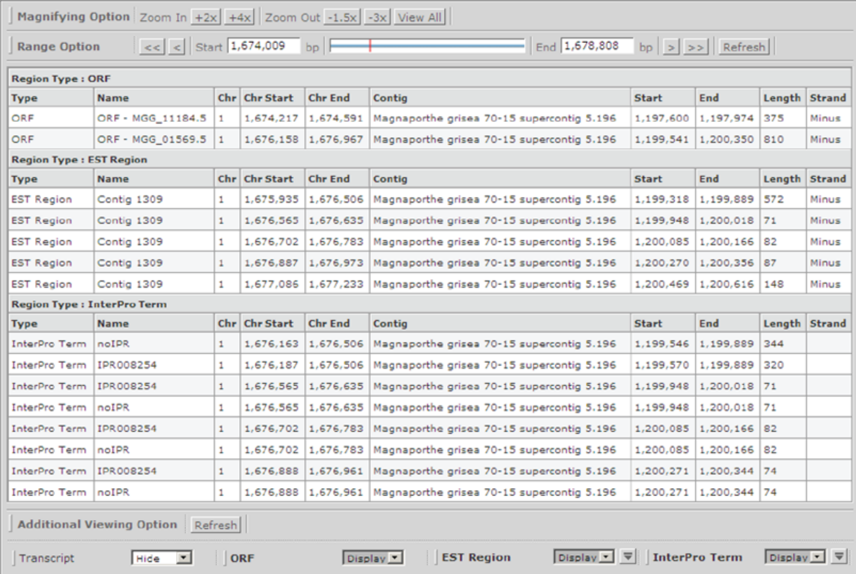
InterPro Term None GC Ratio S.Line

Figure 4. Six different display methods of the genome content and properties via Graphical browser

(A) The graphical browser in SNUGB shows the genome context via three different formats: bar, squash, and pack. At the bottom, ORFs, ESTs, and InterPro domains on chromosome 1 of *M. oryzae* are displayed. (B) Three graphic representations, including graph, line, and Single line (S. line), of the AT-skew, GC-skew, and GC content are shown.

Table browser and Text browser

Although graphical presentation of genomic features helps users view global patterns, the graphical browser does not provide sequences or a list of elements present in a chosen area. To provide such information, we developed two additional tools named as the table browser and the text browser. The table browser provides a list of the names and chromosomal/contig positions of all elements present in a selected region in the csv format, which can be opened using the Excel program (Figure 5A). The text browser provides sequences in a selected region. If ORFs exist in the region, exons and introns are presented using different colors and cases; this function is useful for designing primers and transferring selected sequences to a different data analysis environment (Figure 5B). Additionally, all InterPro domains present on each ORF are displayed as special characters under corresponding sequences so that putative functional domains can be easily recognized at the sequence level. The table and text browser can display sequences up to 50kb.



A

Magnaporthe oryzae 5

Chromosome 1 (Total Length : 8,322,655bp)

Magnifying Option | Zoom In +2x +4x | Zoom Out -1.5x -3x View All

Range Option | << < Start 1,674,009 bp | End 1,678,808 bp > >> Refresh

Type	Name	Chr	Chr Start	Chr End	Contig	Start	End	Length	Strand
Region Type : ORF									
ORF	ORF - MGG_11184.5	1	1,674,217	1,674,591	Magnaporthe grisea 70-15 supercontig 5.196	1,197,600	1,197,974	375	Minus
ORF	ORF - MGG_01569.5	1	1,676,158	1,676,967	Magnaporthe grisea 70-15 supercontig 5.196	1,199,541	1,200,350	810	Minus
Region Type : EST Region									
EST Region	Contig 1309	1	1,675,935	1,676,506	Magnaporthe grisea 70-15 supercontig 5.196	1,199,318	1,199,889	572	Minus
EST Region	Contig 1309	1	1,676,565	1,676,635	Magnaporthe grisea 70-15 supercontig 5.196	1,199,948	1,200,018	71	Minus
EST Region	Contig 1309	1	1,676,702	1,676,783	Magnaporthe grisea 70-15 supercontig 5.196	1,200,085	1,200,166	82	Minus
EST Region	Contig 1309	1	1,676,887	1,676,973	Magnaporthe grisea 70-15 supercontig 5.196	1,200,270	1,200,356	87	Minus
EST Region	Contig 1309	1	1,677,086	1,677,233	Magnaporthe grisea 70-15 supercontig 5.196	1,200,469	1,200,616	148	Minus
Region Type : InterPro Term									
InterPro Term	noIPR	1	1,676,163	1,676,506	Magnaporthe grisea 70-15 supercontig 5.196	1,199,546	1,199,889	344	
InterPro Term	IPR008254	1	1,676,187	1,676,506	Magnaporthe grisea 70-15 supercontig 5.196	1,199,570	1,199,889	320	
InterPro Term	IPR008254	1	1,676,565	1,676,635	Magnaporthe grisea 70-15 supercontig 5.196	1,199,948	1,200,018	71	
InterPro Term	noIPR	1	1,676,565	1,676,635	Magnaporthe grisea 70-15 supercontig 5.196	1,199,948	1,200,018	71	
InterPro Term	IPR008254	1	1,676,702	1,676,783	Magnaporthe grisea 70-15 supercontig 5.196	1,200,085	1,200,166	82	
InterPro Term	noIPR	1	1,676,702	1,676,783	Magnaporthe grisea 70-15 supercontig 5.196	1,200,085	1,200,166	82	
InterPro Term	IPR008254	1	1,676,888	1,676,961	Magnaporthe grisea 70-15 supercontig 5.196	1,200,271	1,200,344	74	
InterPro Term	noIPR	1	1,676,888	1,676,961	Magnaporthe grisea 70-15 supercontig 5.196	1,200,271	1,200,344	74	

Additional Viewing Option | Refresh

Transcript Hide | ORF Display | EST Region Display | InterPro Term Display

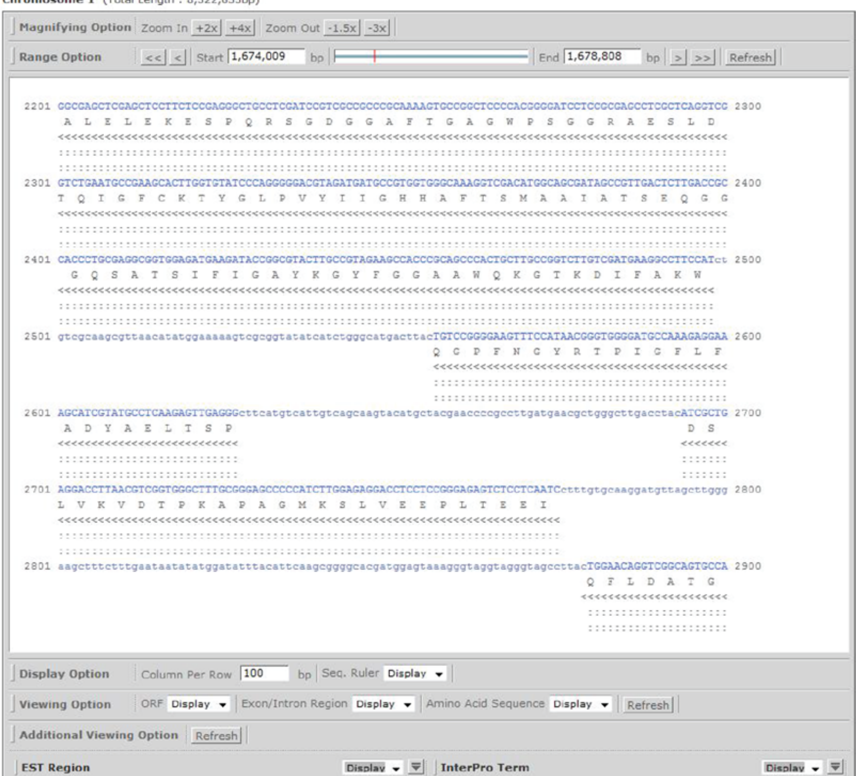


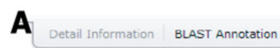
Figure 5. Table and Text browsers

(A) The table browser shows all ORFs, ESTs, and InterPro domains in a selected region as a list. (B) The text browser displays sequences showing exon/intron region as different colors and EST and InterPro domains.

Identification of the putative orthologues of individual fungal proteins kingdom-wide via BLAST and comparison of the genomic contexts and properties of homologous proteins among species via the Session History function

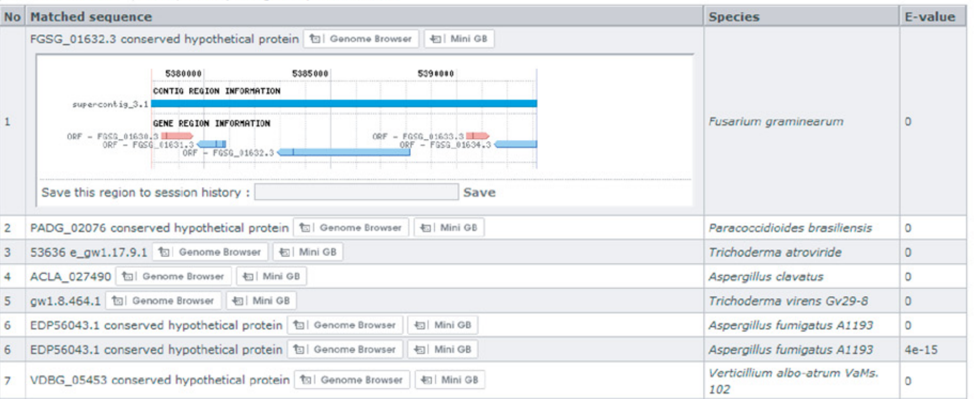
To identify putative orthologues of individual fungal proteins, BLAST searches with each of the 924,343 fungal proteins against all proteins were performed using the e-value of $1e^{-5}$ as the cut-off line. The 'BLAST annotation' tab (see Figure 6A) shows a list of putative orthologues of a chosen gene product in other species with their BLAST e-values. To compare the genomic contexts around the orthologous genes between species or among multiple species, users can store the genomic contexts of the genes using the Session History function, in which the stored genomic contexts can be displayed in one screen (Figure 6B). In each session, other information, such as the GC content and InterPro terms, can also be presented to further support the comparison.

A



BLAST Annotation Result

Dataset	(GenomeBrowser) Fungal BLAST Annotation
Calculated Time	calculated at 2008-08-20 18:42:57
Number of Match	**50** sequences (66 fragments) were matched.



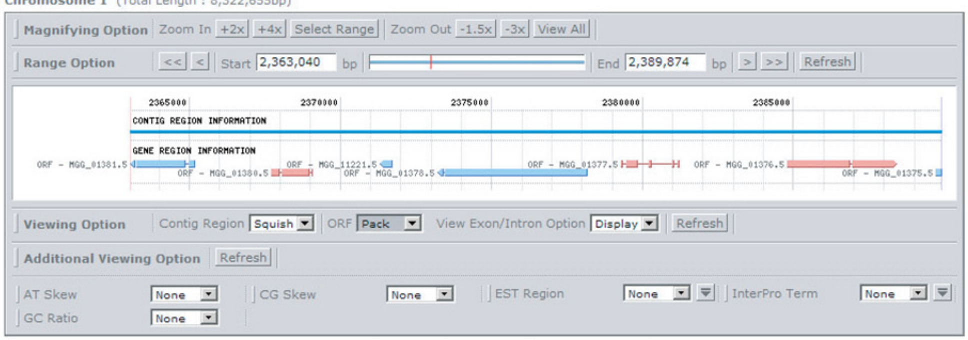
No	Matched sequence	Species	E-value
1	FGSG_01632.3 conserved hypothetical protein [Genome Browser] [Mini GB]	*Fusarium graminearum*	0
2	PADG_02076 conserved hypothetical protein [Genome Browser] [Mini GB]	*Paracoccidioides brasiliensis*	0
3	53636 e_gw1.17.9.1 [Genome Browser] [Mini GB]	*Trichoderma atroviride*	0
4	ACLA_027490 [Genome Browser] [Mini GB]	*Aspergillus clavatus*	0
5	gw1.8.464.1 [Genome Browser] [Mini GB]	*Trichoderma virens Gv29-8*	0
6	EDP56043.1 conserved hypothetical protein [Genome Browser] [Mini GB]	*Aspergillus fumigatus A1193*	0
6	EDP56043.1 conserved hypothetical protein [Genome Browser] [Mini GB]	*Aspergillus fumigatus A1193*	4e-15
7	VDBG_05453 conserved hypothetical protein [Genome Browser] [Mini GB]	*Verticillium albo-atrum VaMs. 102*	0

B

Genome Browser | Text Browser | Table Browser

Magnaporthe oryzae 5

Chromosome 1 (Total Length : 8,322,655bp)

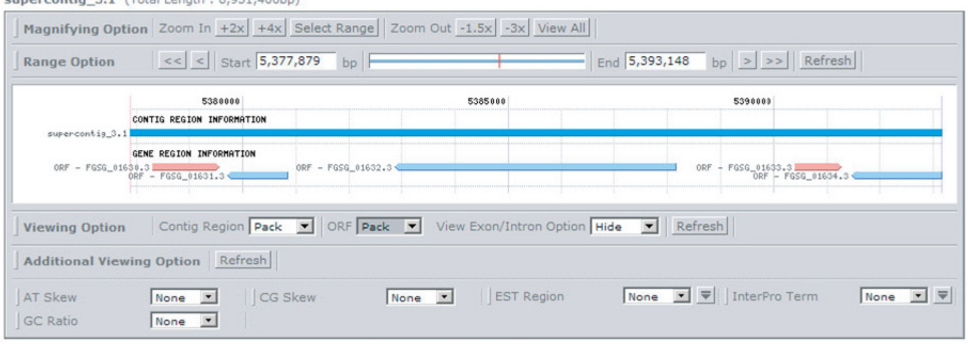


Refresh | Save | Download Image

Genome Browser | Text Browser | Table Browser

Fusarium graminearum 3 PH-1

supercontig_3.1 (Total Length : 8,931,406bp)



Refresh | Save | Download Image

Figure 6. BLAST annotation to catalog homologous proteins

(A) A result of 'BLAST annotation' is shown with the corresponding gene names, species names, and e-values of putative homologs. 'Genome Browser' button after gene name can display the genome context of the selected gene, and 'Mini GB' button will show genome contexts of the selected gene as a smaller size to provide a quick overview, supported by MiniGB. The session can be stored by clicking the save link inside the small SNUGB image. (B) Two independent sessions showing homologs of two genes, MGG_01378.5 and FGSG_01632.3, are shown. Clicking the red button X at the bottom will hide the session.

Additional functionalities of SNUGB

The ‘flexible-range-select’ function allows users to select a chromosomal segment by clicking a mouse at the start site and moving it over the desired segment; the selected area will be displayed as shaded box, and the subsequent click displays an enlarged view of the selected segment (Figure 3A). Through the ‘high-resolution-diagram’ function, users can obtain a high-resolution image (more than 3,000 pixels in width) showing various features on a whole chromosome, such as ORFs, InterPro terms and GC content. This image can be downloaded as image file via both the graphical genome browser and the session-storage function.

CONCLUSION AND DISCUSSION

The SNUGB supports efficient and versatile visualization and utilization of rapidly increasing fungal genome sequence data, as well as those from selected organisms in other kingdoms, to address various types of questions at the genome scale. Properties and features of the archived fungal genomes are available for viewing and comparison in SNUGB. The taxonomy browser helps users easily access the genomes of individual species and provides taxonomic positions of chosen species, and the chromosome map function shows the whole genome of selected species. The graphical browser, table browser, and text browser present a global view of genomic contexts in a selected chromosomal region and support analyses of sequences in the region. The 'BLAST annotation' provides lists of putatively orthologous proteins in the fungal kingdom and facilitates comparison of the genomic contexts of their genes across multiple species. The SNUGB also allows users to manage their own data via the SNUGB web site.

LITERATURE CITED

Adl, S., Simpson, A., Farmer, M., Andersen, R., Anderson, O., Barta, J., Bowser, S., Brugerolle, G., Fensome, R., Fredericq, S., James, T., Karpov, S., Kugrens, P., Krug, J., Lane, C., Lewis, L., Lodge, J., Lynn, D., Mann, D., McCourt, R., Mendoza, L., Moestrup, O., Mozley-Standridge, S., Nerad, T., Shearer, C., Smirnov, A., Spiegel, F., Taylor, M. (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol* **52**: 399 - 451.

AGI (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796 - 815.

Blattner, F., Plunkett, G., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B., Shao, Y. (1997) The complete genome sequence of Escherichia coli K-12. *Science* **277**: 1453 - 1474.

Byrne, K., Wolfe, K. (2005) The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* **15**: 1456 - 1461.

Choi, J., Park, J., Kim, D., Jung, K., Kang, S., Lee, Y.H. (2010) Fungal secretome database: integrated platform for annotation of fungal secretomes. *BMC Genomics* **11**: 105.

Choi, J., Park, J., Jeon, J., Chi, M., Goh, J., Yoo, S., Park, J., Jung, K., Kim, H., Park,

S., Rho, H., Kim, S., Kim, B., Han, S., Kang, S., Lee, Y. (2007) Genome-wide analysis of T-DNA integration into the chromosomes of Magnaporthe oryzae. *Mol Microbiol* **66**: 371 - 382.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B., Johnston, M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**: 71 - 76.

Cornell, M., Alam, I., Soanes, D., Wong, H., Hedeler, C., Paton, N., Rattray, M., Hubbard, S., Talbot, N., Oliver, S. (2007) Comparative genome analysis across a kingdom of eukaryotic organisms: specialization and diversification in the fungi. *Genome Res* **17**: 1809 - 1822.

CSC (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**: 2012 - 2018.

Cuomo, C., Guldener, U., Xu, J., Trail, F., Turgeon, B., Di Pietro, A., Walton, J., Ma, L., Baker, S., Rep, M., Adam, G., Antoniw, J., Baldwin, T., Calvo, S., Chang, Y., Decaprio, D., Gale, L., Gnerre, S., Goswami, R., Hammond-Kosack, K., Harris, L., Hilburn, K., Kennell, J., Kroken, S., Magnuson, J., Mannhaupt, G., Mauceli, E., Mewes, H., Mitterbauer, R., Muehlbauer, G., Munsterkotter, M., Nelson, D., O'Donnell, K., Ouellet, T., Qi, W., Quesneville, H., Roncero, M., Seong, K., Tetko, I., Urban, M., Waalwijk, C., Ward, T., Yao, J., Birren, B., Kistler, H. (2007) The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317**: 1400 - 1402.

Ellegren, H. (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Mol Ecol* **17**: 1629 - 1631.

Fitzpatrick, D., Logue, M., Stajich, J., Butler, G. (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol* **6**: 99.

Flicek, P., Aken, B., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X., Herrero, J., Hubbard, T., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., Searle, S. (2008) Ensembl 2008. *Nucleic Acids Res* **36**: D707 - 714.

Forster, H., Coffey, M., Elwood, H., Sogin, M. (1990) Sequence analysis of the small subunit ribosomal RNAs of three zoosporic fungi and implications for fungal evolution. *Mycologia* **82**: 306 - 312.

Galagan, J., Calvo, S., Cuomo, C., Ma, L., Wortman, J., Batzoglou, S., Lee, S., Basturkmen, M., Spevak, C., Clutterbuck, J., Kapitonov, V., Jurka, J., Scazzocchio, C., Farman, M., Butler, J., Purcell, S., Harris, S., Braus, G.,

Draht, O., Busch, S., D'Enfert, C., Bouchier, C., Goldman, G., Bell-Pedersen, D., Griffiths-Jones, S., Doonan, J., Yu, J., Vienken, K., Pain, A., Freitag, M., Selker, E., Archer, D., Penalva, M., Oakley, B., Momany, M., Tanaka, T., Kumagai, T., Asai, K., Machida, M., Nierman, W., Denning, D., Caddick, M., Hynes, M., Paoletti, M., Fischer, R., Miller, B., Dyer, P., Sachs, M., Osmani, S., Birren, B. (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* **438**: 1105 - 1115.

Gregory, T. (2004) Insertion-deletion biases and the evolution of genome size. *Gene* **324**: 15 - 34.

Hedeler, C., Wong, H., Cornell, M., Alam, I., Soanes, D., Rattray, M., Hubbard, S., Talbot, N., Oliver, S., Paton, N. (2007) e-Fungi: a data resource for comparative analysis of fungal genomes. *BMC Genomics* **8**: 426.

Hibbett, D., Binder, M., Bischoff, J., Blackwell, M., Cannon, P., Eriksson, O., Huhndorf, S., James, T., Kirk, P., Lucking, R., Thorsten Lumbsch, H., Lutzoni, F., Matheny, P., McLaughlin, D., Powell, M., Redhead, S., Schoch, C., Spatafora, J., Stalpers, J., Vilgalys, R., Aime, M., Aptroot, A., Bauer, R., Begerow, D., Benny, G., Castlebury, L., Crous, P., Dai, Y., Gams, W., Geiser, D., Griffith, G., Gueidan, C., Hawksworth, D., Hestmark, G., Hosaka, K., Humber, R., Hyde, K., Ironside, J., Koljalg, U., Kurtzman, C., Larsson, K., Lichtwardt, R., Longcore, J., Miadlikowska, J., Miller, A., Moncalvo, J., Mozley-Standridge, S., Oberwinkler, F., Parmasto, E., Reeb, V., Rogers, J., Roux, C., Ryvarden, L., Sampaio, J., Schussler, A., Sugiyama, J., Thorn, R.,

Tibell, L., Untereiner, W., Walker, C., Wang, Z., Weir, A., Weiss, M., White, M., Winka, K., Yao, Y., Zhang, N. (2007) A higher-level phylogenetic classification of the Fungi. *Mycol Res* **111**: 509 - 547.

Jeon, J., Park, S., Chi, M., Choi, J., Park, J., Rho, H., Kim, S., Goh, J., Yoo, S., Choi, J., Park, J., Yi, M., Yang, S., Kwon, M., Han, S., Kim, B., Khang, C., Park, B., Lim, S., Jung, K., Kong, S., Karunakaran, M., Oh, H., Kim, H., Kim, S., Park, J., Kang, S., Choi, W., Kang, S., Lee, Y. (2007) Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. *Nat Genet* **39**: 561 - 565.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241 - 254.

Kuhn, R., Karolchik, D., Zweig, A., Trumbower, H., Thomas, D., Thakkapallayil, A., Sugnet, C., Stanke, M., Smith, K., Siepel, A., Rosenbloom, K., Rhead, B., Raney, B., Pohl, A., Pedersen, J., Hsu, F., Hinrichs, A., Harte, R., Diekhans, M., Clawson, H., Bejerano, G., Barber, G., Baertsch, R., Haussler, D., Kent, W. (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res* **35**: D668 - 673.

Lee, W., Park, J., Choi, J., Jung, K., Park, B., Kim, D., Lee, J., Ahn, K., Song, W., Kang, S., Lee, Y.H., Lee, S. (2009) IMGD: an integrated platform supporting comparative genomics and phylogenetics of insect mitochondrial genomes. *BMC Genomics* **10**: 148.

Loftus, B., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I., Fraser, J., Allen, J., Bosdet, I., Brent, M., Chiu, R., Doering, T., Donlin, M., D'Souza, C., Fox, D., Grinberg, V., Fu, J., Fukushima, M., Haas, B., Huang, J., Janbon, G., Jones, S., Koo, H., Krzywinski, M., Kwon-Chung, J., Lengeler, K., Maiti, R., Marra, M., Marra, R., Mathewson, C., Mitchell, T., Pertea, M., Riggs, F., Salzberg, S., Schein, J., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C., Suh, B., Tenney, A., Utterback, T., Wickes, B., Wortman, J., Wye, N., Kronstad, J., Lodge, J., Heitman, J., Davis, R., Fraser, C., Hyman, R. (2005) The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science* **307**: 1321 - 1324.

Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A., Orchard, S., Orengo, C., Petryszak, R., Selengut, J., Sigrist, C., Thomas, P., Valentin, F., Wilson, D., Wu, C., Yeats, C. (2007) New developments in the InterPro database. *Nucleic Acids Res* **35**: D224 - 228.

Okoniewski, M., Yates, T., Dibben, S., Miller, C. (2007) An annotation infrastructure for the analysis and interpretation of Affymetrix exon array data. *Genome*

Biol **8**: R79.

Park, J., Lee, S., Choi, J., Ahn, K., Park, B., Park, J., Kang, S., Lee, Y. (2008a) Fungal Cytochrome P450 Database. *BMC Genomics* **9**: 402.

Park, J., Kim, H., Kim, S., Kong, S., Park, J., Kim, S., Han, H., Park, B., Jung, K., Lee, Y.-H.-. (2006) A comparative genome-wide analysis of GATA transcription factors in fungi. *Genomics & Informatics* **4**: 147 - 160.

Park, J., Park, J., Jang, S., Kim, S., Kong, S., Choi, J., Ahn, K., Kim, J., Lee, S., Kim, S., Park, B., Jung, K., Kim, S., Kang, S., Lee, Y. (2008b) FTFD: an informatics pipeline supporting phylogenomic analysis of fungal transcription factors. *Bioinformatics* **24**: 1024 - 1025.

Park, J., Park, B., Jung, K., Jang, S., Yu, K., Choi, J., Kong, S., Park, J., Kim, S., Kim, H., Kim, S., Kim, J., Blair, J., Lee, K., Kang, S., Lee, Y. (2008c) CFGP: a web-based, comparative fungal genomics platform. *Nucleic Acids Res* **36**: D562 - 571.

Park, J., Park, B., Veeraraghavan, N., Jung, K., Lee, Y., Blair, J., Geiser, D., Isard, S., Mansfield, M., Nikolaeva, E., Park, S., Russo, J., Kim, S., Greene, M., Ivors, K., Balci, Y., Peiman, M., Erwin, D., Coffey, M., Rossman, A., Farr, D., Cline, E., Crunwald, N., Luster, D., Schrandt, J., Martin, F., Ribeiro, O., Makalowska, I., Kang, S. (2008d) Phytophthora Database: A Forensic Database Supporting the Identification and Monitoring of Phytophthora. *Plant Dis* **92**: 966 - 972.

Scannell, D., Byrne, K., Gordon, J., Wong, S., Wolfe, K. (2006) Multiple rounds of

speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**: 341 - 345.

Shiu, S., Borevitz, J. (2008) The next generation of microarray research: applications in evolutionary and ecological genomics. *Heredity* **100**: 141 - 149.

Soanes, D., Alam, I., Cornell, M., Wong, H., Hedeler, C., Paton, N., Rattray, M., Hubbard, S., Oliver, S., Talbot, N. (2008) Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *PLoS ONE* **3**: e2300.

Stanke, M., Steinkamp, R., Waack, S., Morgenstern, B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* **32**: W309 - 312.

Stein, L., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J., Harris, T., Arva, A., Lewis, S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res* **12**: 1599 - 1610.

Thomas, C. (1971) The genetic organization of chromosomes. *Annu Rev Genet* **5**: 237 - 256.

Wheeler, D., Barrett, T., Benson, D., Bryant, S., Canese, K., Chetvernin, V., Church, D., DiCuccio, M., Edgar, R., Federhen, S., Geer, L., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D., Madden, T., Maglott, D., Ostell, J., Miller, V., Pruitt, K., Schuler, G., Sequeira, E., Sherry, S., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R., Tatusova, T., Wagner, L., Yaschenko, E. (2007) Database resources of the National Center for Biotechnology Information.

Nucleic Acids Res **35**: D5 - 12.

Wolfe, K. (2006) Comparative genomics and genome evolution in yeasts. *Philos Trans R Soc Lond B Biol Sci* **361**: 403 - 412.

Xi, H., Park, J., Ding, G., Lee, Y.H., Li, Y. (2009) SysPIMP: the web-based systematical platform for identifying human disease-related mutated sequences from mass spectrometry. *Nucleic acids research* **37**: D913-920

Xu, J., Saunders, C., Hu, P., Grant, R., Boekhout, T., Kuramae, E., Kronstad, J., Deangelis, Y., Reeder, N., Johnstone, K., Leland, M., Fieno, A., Begley, W., Sun, Y., Lacey, M., Chaudhary, T., Keough, T., Chu, L., Sears, R., Yuan, B., Dawson, T. (2007) Dandruff-associated Malassezia genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens. *Proc Natl Acad Sci USA* **104**: 18730 - 18735.

Yates, T., Okoniewski, M., Miller, C. (2008) X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Res* **36**: D780 - 786.

CHAPTER 2

Laboratory Information Management System for the fungal pathogenomics

ABSTRACT

Rapidly accumulating genome sequences have accelerated the generation of diverse types of large data sets. However, comparison of such datasets within individual communities are often quite challenging, because they typically have been generated via the use of laboratory-specific protocols and data acquisition formats rather than a set of standards agreed upon by the whole community. To stimulate the adoption of standardized, shared experimental protocols and data acquisition formats by fungal research communities, a Laboratory Information Management System (LIMS) was developed using *Magnaporthe oryzae*, a plant pathogenic fungus, as a model. This LIMS, entitled the LIMS for *M. oryzae* (LIMSMO), provides a detailed guideline for key phenotype assays and standardized formats for data acquisition, thus facilitating quantitative comparison of newly acquired data with those previously acquired. All phenotypic data derived from functional studies of individual genes are linked to the corresponding gene locus information, which allows further bioinformatics analyses of the gene of interest through the use of Comparative Fungal Genomics Platform (http://cfgp.riceblast.snu.ac.kr). The LIMSMO can be easily modified to support systematic pathogenomics research on other fungal pathogens.

INTRODUCTION

Since the release of yeast genome sequence in 1996 (Goffeau *et al.*, 1996), more than 100 species have been published (Choi *et al.*, 2013) with many more species currently being sequenced. The amount of experimental results has been growing exponentially in large part due to subsequent functional genomic analyses. However, a majority of resulting datasets are not easily comparable between laboratories, because data acquisition often has been conducted using laborabory-specific experimental protocols and data acquisition formats. To ensure efficient and comprehensive investigations of complex biological problems by integrating rapidly accumulating experimental data from multiple laboratories, use of globally accepted standardized experimental protocols and data acquisition formats is critical. Otherwise community work will get fragmented and become inefficient.

The Laboratory Information Management System (LIMS) is a package of computer software used for the management of experimental samples, resulting data, workflow of laboratory members, instruments and protocols. It can also manage other laboratory functions such as invoicing, sample preparation, and workflow automation (Robinson, 1983). Various LIMSs have been developed to support: a) information gathering within laboratory networks, b) supervision of laboratory work processes, c) identification and resolution of major bottlenecks in workflows, d) calculate and maintain processing and handling times on chemical reactions, e) documentation and review of work progress and resulting data via online databases and f) remote control

of tasks and their progress (Skobelev *et al.*, 2011).

For example, to support the integration, management and analysis of vast amounts of microarray data, LIMS for BioArray Software Environment (BASE) was developed. The BASE supports the integration of biomaterial information, raw experimental images and extracted data, and also provides tools for data transformation, viewing and analysis (Saal *et al.*, 2002). Furthermore, LIMSs also have been developed to manage protein-protein interaction data, a web-based tool supporting local protein sequence annotation, protein database searching and sequence management (Droit *et al.*, 2007).

An enormous amount of genetic and physical map data derived from the maize mapping project (MMP) is also managed by a LIMS, which provides iMap or cMap software to verify SSR, RFLP, SNP and InDel markers (Sanchez-Villeda *et al.*, 2003). Another LIMS platform functions to support candidate gene mutation screening by assisting communications with other lab instrument (LightScanner or BioRobot), tracking samples and laboratory information, and gathering data at every step of screening workflow (Voegele *et al.*, 2007). To support functional genomic analysis via the use of high throughput generation and screening of random insertional mutants of *Magnaporthe grisea*, PACLIMS (Phenotype Assay Component LIMS), a system related to MGOS (Kour *et al.*, 2012), was built (Donofrio *et al.*, 2005; Soderlund *et al.*, 2006).

The LIMS also has been applied in various ways for managing and querying phenotypic data in model organisms. To support the characterization of more than

5,000 mutants of *Arabidopsis* through reverse genetics phenotypic screening, a LIMS was implemented (Lu *et al.*, 2011). To manage materials and work associated with a large-scale insertional mutagenesis project of rice, including seeds, transformation process, plant/progeny and phenotype information, the Rice Gene Machine Information Management System (RGMIMS) was developed (Henry *et al.*, 2008). In the field of microbial phenomics, the *Saccharomyces* Genome Database (SGD) is supported by a curation system based on a GO term controlled-vocabulary system to facilitate querying and comparison of mutant phenotypes (Costanzo *et al.*, 2009).

There exists no LIMS platform for filamentous fungi, which include important plant pathogens like *Magnaporthe oryzae*. *M. oryzae* is an ascomycete fungus that causes rice blast, one of the most destructive diseases throughout the world. Rice blast disease causes a yield loss enough for feeding about 60 million people in a year (Talbot, 2003). *M. oryzae* is an important model organism for studying molecular plant-pathogen interactions due to its genetic tractability and the availability of genome sequences of both rice and *M. oryzae* (Dean *et al.*, 2005; Ebbole, 2007). Functions of *M. oryzae* genes including transcription factor (Kim *et al.*, 2009) and non-transcription factor (Chi *et al.*, 2009) have been extensively studied by making gene knock-out mutants, resulting in many experimental results. Large scale genome-wide forward genetics approaches via the use of transformation-mediated insertional mutagenesis also have been used to generate a large number of insertion mutants (Balhadere *et al.*, 1999; Chen *et al.*, 2011; Jeon *et al.*, 2007). High-throughput phenotype screening pipelines and computerized database were developed to identify

interesting genes and archiving resulting data, respectively (Jeon *et al.*, 2007). However, most available experimental data are difficult to compare, because no globally accepted standardized experimental protocols and data acquisition formats have been used. Although much progress has been made in plant and animal genomics studies via the use of LIMS, a comparable effort to effectively manage functional genomics data from filamentous fungi is urgently needed.

To address this need, we developed LIMS for *M. oryzae* (LIMSMO) to provide not only an integrated management system for results derived from targeted gene knock-out, but also standardized methods including a guideline for phenotype assays and data acquisition formats. In addition, we also linked phenotype data in LIMSGO to gene sequence-associated information in CFGP (http://cfgp.snu.ac.kr), a genome data and data analysis tool warehouse (Choi *et al.*, 2013), through the locus id of individual genes. This system manages mutants via their locus numbers, experimental phenotypic data through a task list, lab standard protocols, and relevant literatures. Sharing all standardized experimental results between members of the global *M. oryzae* research community is supported by LIMSMO. The putative function of unknown genes can be more easily speculated by enabling systematic comparison of mutant phenotypic data with pre-existing data in the user-friendly web-based LIMSMO database.

MATERIALS AND METHODS

I. Development of standardized phenotype assays based on published papers

The PubMed database (http://www.ncbi.nlm.nih.gov/pubmed/) was searched using the keywords “Magnaporthe”, “Rice blast” and “Gene Characterization” to identify relevant papers. We picked up papers that they have sequence information about characterized genes. If they did not provide the locus number but only gene sequences, the sequences were mapped to *Magnaporthe oryzae* ver. 6 genome by BLAST search to find locus number. Finally, the locus tagged papers were categorized by phenotype assay.

II. Server operation

The LIMSMO software was based on GNU/Linux operation system (Fedora release 8) and was developed using PHP (version 5.2.6) language. Web-service is running on Apache (version 2.2.9) web server.

III. Database composition

All experimental data and other information table is stored in a relational database (MySQL; version 5.0.88) and managed by its DBMS. Remote servers for the database were physically separated from the web server, and these servers were linked by an

internal gigabit network using hard wired setup.

IV. External program

For the statistic calculation for experimental results, we used R (version 2.8.0) statistical computing language. Tukey's HSD Post-Hoc calculation was derived from agricolae package (http://cran.r-project.org/web/packages/agricolae/index.html) provided by R program automatically linked with PHP by shell script.

V. Web system for LIMSMO

LIMSMO is available at http://lims.riceblast.snu.ac.kr to allow anyone who studies *M. oryzae*. Details about how to use LIMSMO are shown in Figures S2-S10 in supplement data. This system can be accessed using any web browser in any operating system that can present HTML5 with CSS3 (Internet Explorer 9 or higher, Safari 5 or higher, Chrome 10 or higher, Firefox 4 or higher).

RESULTS

I. Development of standardized protocols

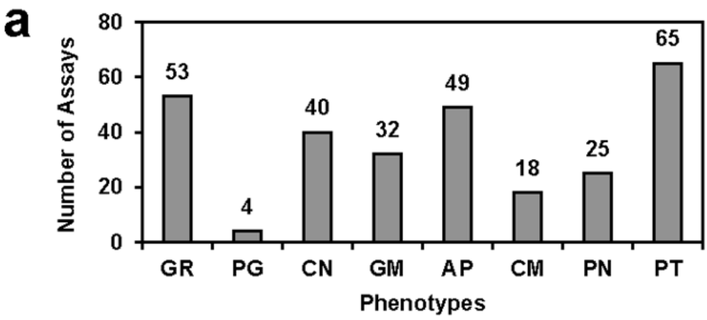
We selected 126 papers describing functional characterization of *M. oryzae* genes published from 1987 to 2011 by searching the PubMed (http://pubmed.org) database (Figure S1 in supplement data). A group of 57 papers having sequence information or locus number were used in phenotype of mutant characterization. Subsequently, we collected 72 genes from selected papers, but 4 genes from the papers were conducted as a duplicate (Table S1 in supplement data).

Experimental phenotype assays were categorized included growth rate, mycelial pigmentation, conidiation, conidial germination, appressorium formation, conidial morphology, appressorium penetration and pathogenicity. For each category, the characterized genes were counted as the number of task for an experimental method and different methods were counted as a separated task (Table S1 in supplement data). The most experimented assays in ascending order are pathogenicity (65 assays), growth rate (53 assays), appressorium formation (49 assays), conidiation (40 assays), germination (32 assays), appressorium penetration (25 assays), conidial morphology (18 assays) and pigmentation (4 assays) (Figure 1a). This result shows inequality of assayed tasks for each phenotype assay.

We also organized and checked over 300 materials and methods for the phenotype assay (Table S2 in supplement data). Every categorized assay section was subdivided into protocol types including the media types, treatment methods, measurement

methods and so forth. In this data, the experiment on the growth rate assay used 113 methods and appressorium formation used57 methods. Counting in ascending order, pathogenicity (49 methods), germination (36 methods), conidiation (27 methods), appressorium penetration (15 methods), pigmentation (10 methods) and conidial morphology (1 method) assays were experimented in such result. Thus, we realized the number of methods used for each phenotype assay were uneven (Figure 1b).

Consequentially, biased research in assay type and irregular protocols cause fragmented result and hardship for sharing data. To complement this bias, our LIMS system provides guidelines for experimental assay and acquiring result in standardized format. Refer to the most used protocols in result of collection above, we made global standardized experimental protocols containing *M. grisea* phenotype assay. The protocols not only contain type of treatment, reagent or culture media, but also give the particulars of measurement method or observation time in detail (fully described in section of methods). (Figure S2 in supplement data).



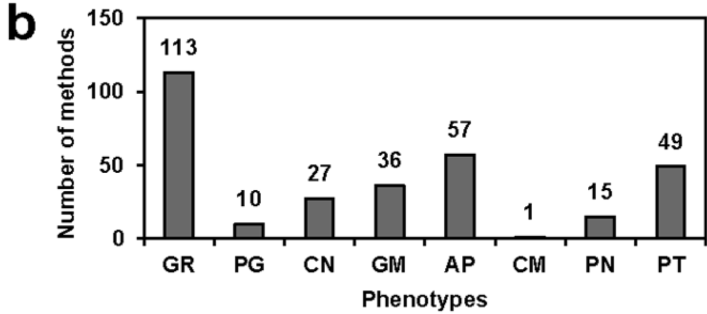


Figure 1. The type and frequency of phenotype assays used in in published papers

(a) Number of assay frequencies for each phenotype from collected papers. **(b)** Number of methods used for each phenotype assay from collected papers. Abbreviation means GR:growth rate of fungal mycelium, PG:mycelial pigmentation, CN:ability to produce conidia, GM:conidial ermination, AP:appressorium formation, CM:conidial morphology, PN:appressorium penetration to host surface, PT:pathogenicity on host.

II. System structure and overall function of LIMSMO

Next step, we developed a computerized database and management system called Laboratory Information Management System for *M. oryzae* for storing and managing experimental data. LIMSMO is a multi-tier client-server application and can be subdivided into different sections according to their functionally defined responsibilities (Figure 2). In the LIMSMO system, the presentation tier consists of a web-interface, using jQuery JavaScript library, CSS3 and HTML5 as GUI components and APM (Apache, PHP, MySQL) web service, which allow access to inner application logic at the business tier. Accordingly, End-user interface of LIMSMO is web-based application, so it does not require special client-side installation. However, it offers users the opportunity to see analytical results in real time from any web browsers and from any operation systems. The LIMSMO system also supports unlimited user connections in simultaneous time.

The business tier is formed by application logic and functional modules, which retrieves and acquires the experimental results dataset by communicating with the database tier. All business logic can be defined once within the business layer and then shared by any number of components within the presentation layer. Therefore any changes to business rules can be made in one place and be instantly available throughout the whole application. Creating a new project of phenotype assay and storing dataset of phenotype assay result which follows laboratory workflow are the first section of application logic in business tier. The second part of the application logic is protocol section. All protocols used in phenotype assay or newly created are

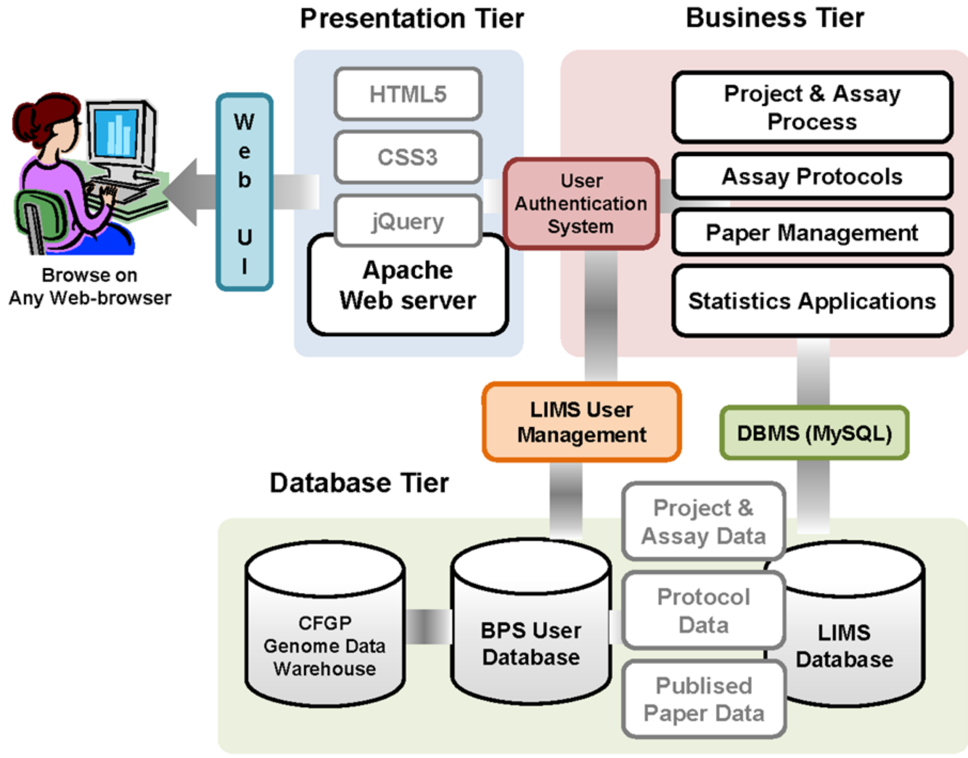


Figure 2. The software architecture of LIMSMO

This LIMS system is based on a three-tier architecture, including presentation, business and database. The presentation tier consists of a web-interface with GUI components and APM (Apache, PHP, MySQL) web service. The business tier contains application logic and functional modules and retrieves experimental results by communicating with the database tier. External programs for statistical analyses are also implemented and managed by the application logic. All application logic in the business tier is protected by an authentication logic entitled the passport system, which controls user authorization and manages user information. The database tier

consists of the DBMS (database management system) and databases. Remote servers of database system are physically separated, and the business tier is bound by network using hard wired setup.

saved and managed by this section. Published research papers about assay and its related locus information are also collected and interconnected by our system. External programs for statistical calculation are also implemented and managed by application logic. All experimental data which has repetition are statistically analyzed by external program, and its statistical result is stored in database tier. Moreover, application logic of interconnection between locus information in phenotype assay and genotype information in CFGP (http://cfgp.snu.ac.kr), genome data warehouse, is important part of phenotype/genotype linkage. All application logic in business layer is protected by authentication logic called as passport system, which enforces user authorization and manages user information.

The database tier consists of the DBMS (database management system) and the database. Application logic connected to the database all has LIMS instrument data and user information and CFGP database has genomic information. Each connection is secured by authentication logic and transferring of raw data is encrypted. Database managing system is MySQL software of DBMS and data retrieving software is library in PHPs. Remote servers of database system are physically separated and the business tier is bound by internal gigabit network using hard wired setup. Following advantages of multi-tier architecture, separated data access layer provides more than one applications are accessing the same database and vice versa. If any changes occur in the database or any code change is requrired due to an update issue or to fix bugs the changes can be done in that specific component alone.

Database scheme of LIMSMO system is constructed by relational databases (Figure

3). Following principle of relational databases, each table is connected by primary and foreign keys. Tables are grouped by its following function and stored information for mutant phenotype assay process based on experimental design in laboratory. Basic project information about phenotype assay and related paper information are stored to each table, and these are grouped by project setup phase. In the stage of making mutant and acquiring phenotype assay data, assayed locus information that belongs to a project is stored locus information table. The acquired results and used protocols are separated by each assay and stored to result information table and protocol information table. In the aspect of structural view, project information table, locus assay information and tables for tasks of eight phenotype assays having protocol information are hierarchically organized and connected.. For managing efficiency, all protocol information tables of each task are grouped and managed by integrated protocol information table. For the user convenience and tracking ability, all activities performed by researchers are logged into the history information table and all locus information in locus of assay and paper database for LIMSMO is linked to CFGP by sequence id.

The web based end-user interface of LIMSMO system is composed of four large components separated by its functionality: 1) public data section, 2) my data section, 3) protocol section, 4) publication section (Figure 4a). In the public data section, the assay for phenotype experimental result acquired for making guideline of standard protocols in 67 loci are shown in detail. The other experimental data derived from personal phenotype assay project which is released to public by user will be shown in

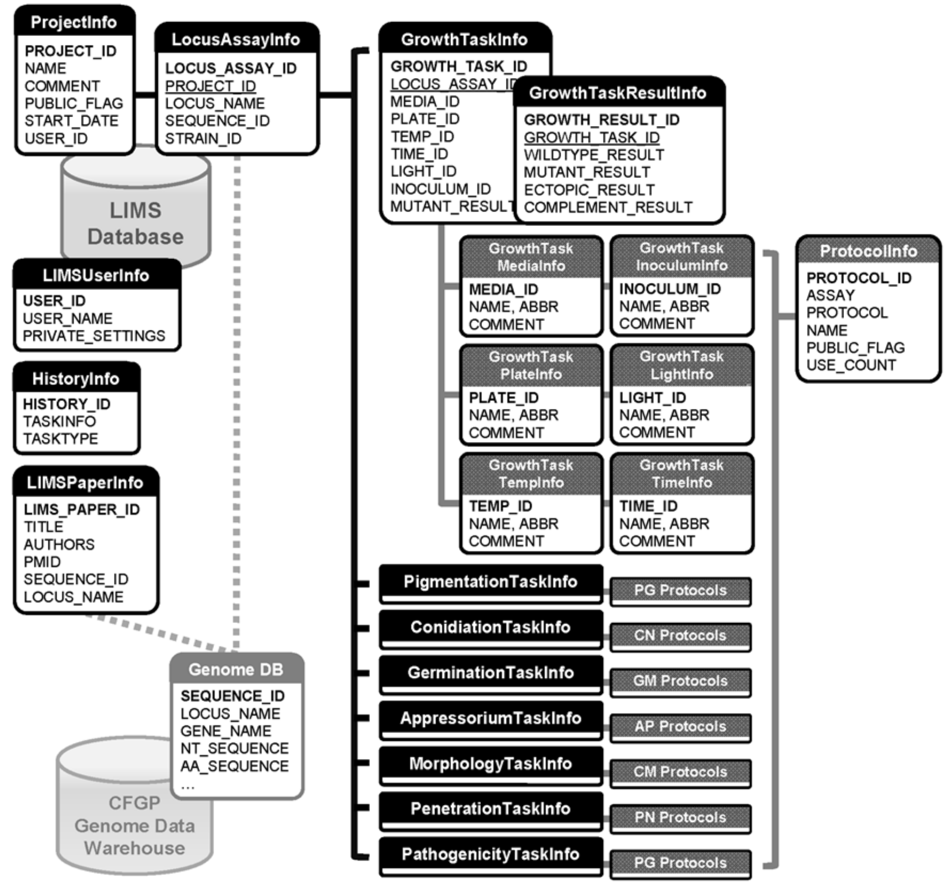


Figure 3. Database scheme of LIMSMO

The bold characters indicate primary key of field name in each table. The underlined characters denote foreign keys related to other tables. Tables for tasks of eight phenotype assays have protocol information tables. All protocol information tables of each task are managed by integrated protocol information table. Activities that occurred by researchers are logged into history information table. All loci informations in locus of assay and paper database for LIMSMO are linked to CFGP, a genome data warehouse, by sequence id.

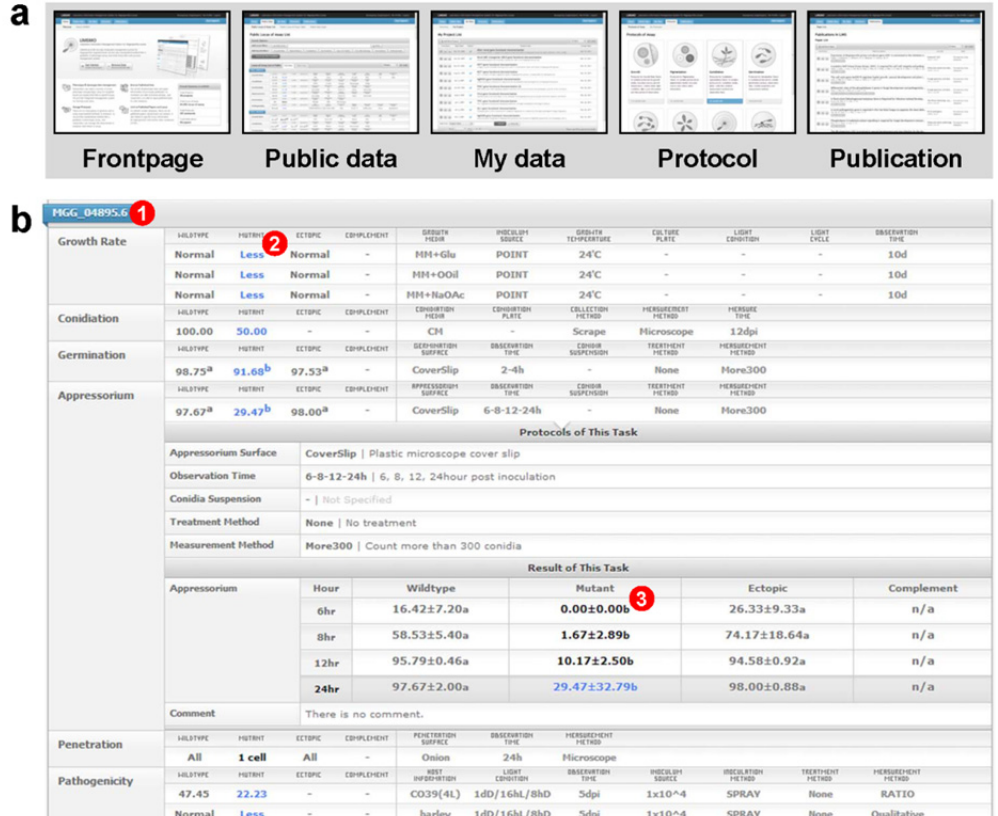


Figure 4. Web based end-user interface of LIMSMO

(a) The LIMSMO is composed of the following four large components separated by their functionality: 1) public data section, 2) my data section, 3) protocol section, and 4) publication section. (b) Users are able to search data of interest by using the filter mask of locus and protocols. The main features of public data section are: 1) Locus number is a link to access the CFGP genome data warehouse; 2) Derived assayed locus and phenotype data from searching mask, results are represented by colors. Red color means over/large status, and blue color means less/small status; and 3) The

superscript alphabet on result represents the group of Tukey's HSD Post-Hoc calculation. (Significance level is 95%)

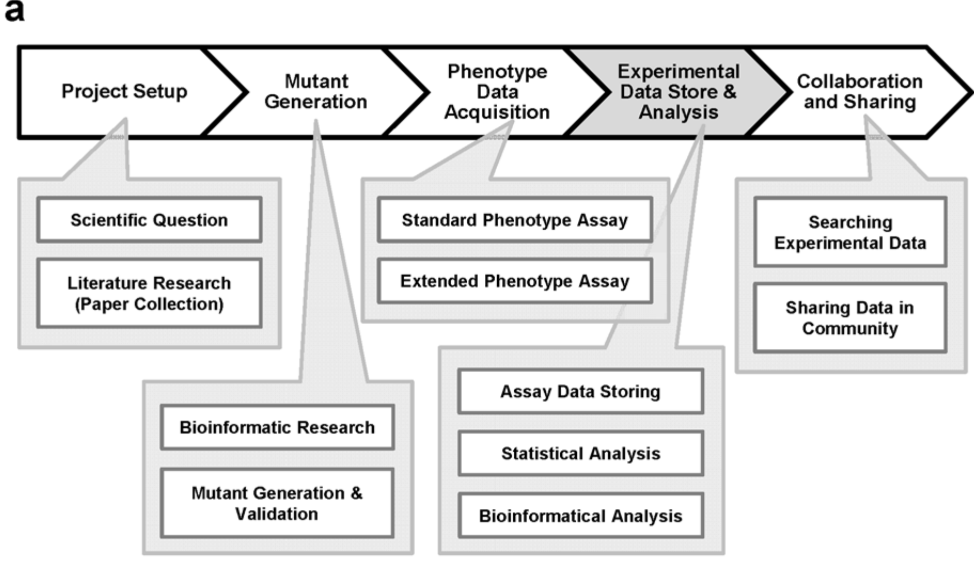
this section (Figure 4b). If users want to make personal dataset of phenotype assay, they can create projects and tasks of each phenotype experiment assigned to this project in “My data” section. Users can also track their activities such as input of new data or modifying data at “Overview” menu in this section. The standardized protocols which already established standard operation procedures or modified and optimized for the specific biological experiment by each user are defined as a procedure of phenotype experimental assay is stored in “Protocol” section. Information of published papers in relation to specific locus is also stored in “Publication” section. Anybody can browse and search papers by locus name or paper information such as abstract or authors. Users can also reach genotype information on CFGP through the linkage of locus information in LIMSMO system that shown anywhere in webpage.

III. Management of experimental data via the workflow-driven experimental design of LIMSMO

Acquired experimental results and protocols used in different laboratory instruments used for making guideline of standard protocols are entered into LIMSMO system by following manners. A series of process about functional characterization of specific gene in *M. oryzae* is divided into five parts of experimental step in workflow: 1) project setup, 2) mutant generation, 3) phenotype data acquisition, 4) experimental data store & analysis and 5) collaboration and data sharing (Figure 5a).

The first step of project starts with a scientific question about determining the function of specific genes in *M. oryzae*. The publication information and related locus for linkage to genotype information are entered into “Publication” section in LIMSMO system. The creation and setting up a new project for one research paper in “My data” section is the start of managing result and methods in phenotype assay process. Each locus entered into project has the information about locus name or gene name and sequence id which are linked to genome database in our data warehouse CFGP.

In the phenotype data acquisition step, eight types of standard phenotype assay and extended phenotype assay such as checking expression of some gene in specific condition or other in-depth phenotype assay are tested by standardized assay protocol. Acquired results of phenotype assays are stored to database in standardized format. Eight types of standard phenotype assay data are entered into each locus in the locus



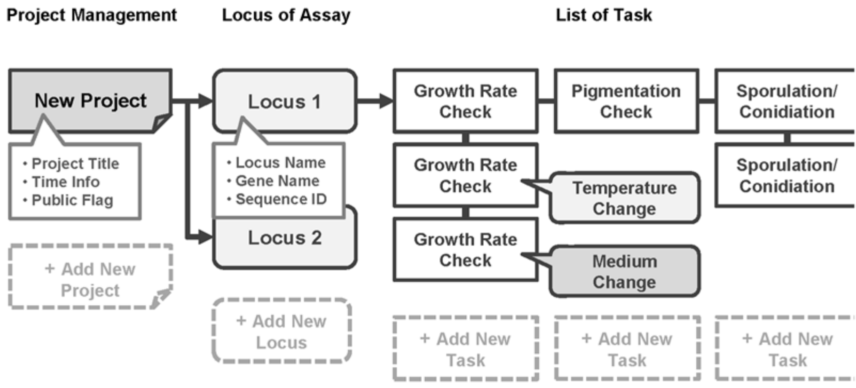


Figure 5. Processes of data acquisition and archiving

(a) A series of process about targeted knock-out assay. Literature research about some scientific questions is the first step in this project. In the mutant generation step, researchers generate some KO mutants based on bioinformatics research. Acquired

result of phenotype assay is stored and analyzed statistically or bioinformatically. In the final step, all results that derived from the previous steps are share to community members. **(b)** Process of phenotype assay in LIMSMO. Create and set up a project is the first step of managing for phenotype assay process. A project consists of project title, description about the project, start-end date and public flag for opening the project to the public. One project is able to have multiple locus of assay. Each locus has the information about locus name or gene name and sequence id that link to genome database in our data warehouse CFGP. Inside a locus of assay, there are series of assay tasks for each phenotype for mutant characterization assay. If you want to change the temperature condition for growth rate check or growth medium, you can register a new task for growth rate check assay task.

of assay subsection. Inside a locus of assay, there are series of assay tasks for each phenotype for mutant characterization assay. If you users want to change the temperature condition for growth rate or growth medium, they can register a new task for growth rate assay task. In this step, materials and methods used for this phenotype experiment are newly entered or selected from previously used protocols that are associated with formerly defined protocol types. These steps of entering experimental result for assay task are supported by wizard-based input user-interface. These stored experimental data linked by locus number in CFGP can be analyzed statistically or bioinformatically to get further information (Figure 5b).

In the final step, all of the result data derived from previous step are able to be searched and shared by peers or community members who want specific conditions.,. In our system, users can export and share experimental result in “Public Data” section for mutual benefits. Input result data using “Public Data Input” wizard when users want to input large scale experimental data by sequential method and independent from time schedule.

Newly entered or modified experimental results and protocols in project are recorded with time stamp, thus all activities are retraceable to all users. It is important features in LIMSMO system, because researchers need to find out what experiments are already completed and enables to get information from old notes. Consequently, LIMSMO provides search filters which admit locus name and assay condition based searching in the database stored experimental results and methods at “Public Locus of Assay List” subsection in “Public Data” section if the public flag of phenotype assay

project is not private but public.

To collect a locus of interest for specific purpose, users are able to combine any type of experimental method and result in phenotype assay by and/or logic. Then, experimental result and protocols of locus are filtered by above filter mask. If users are specifically interested in group of locus or experimentally related locus, users can make favorite which is a group of locus stored in database. The stored favorite is shared to everyone in the LIMSMO system. The phenotype assay results are represented by colors. Red color means over/large status, and blue color means less/small status. The superscript alphabet on result represents the group of Tukey's HSD Post-Hoc calculation with 95% of significance level.

IV. Case study

To validate and test the LIMSMO platform, we followed a direction of phenotype data acquisition and data retrieval process by inputting a series of experimental data generated in our lab. A total of eight genes (MoHOX1 to MoHOX8) encoding putative homeobox transcription factors (TFs) were identified from the *M. oryzae* genome. Knockout mutants for each MoHOX gene were obtained via homology-dependent gene replacement (Kim *et al.*, 2009). The protocols used for characterizing these mutants, and resulting experimental result are good testers for the LIMSMO platform.

Registration of published papers to the "Publication section" is the first step of LIMSMO process. Mentioned and assayed eight genes (MoHOX1; MGG_04853.6, MoHOX2; MGG_00184.6, MoHOX3; MGG_01730.6, MoHOX4; MGG_06285.6, MoHOX5; MGG_07437.6, MoHOX6; MGG_11712.6, MoHOX7; MGG_12865.6, MoHOX8; MGG_12958.6) were also linked to paper information. Setup a new project for generation of MoHOX deletion mutants and observation of their phenotypes are the base process for input assay information. All of the MoHOX deletion mutants were performed six types of standard phenotype assays (growth rate, conidiation, conidial germination, appressorium formation, conidial morphology and host pathogenicity). Each type of phenotype assay has only 1 task, so total 48 tasks of assay were stored in MoHOX deletion mutant project. Assay protocols were shared with every deletion mutants because of the same phenotype experiment were performed to all. Except the conidial morphology assay, as all experiments were

performed and stored in LIMSMO result database with three replicates. Lastly, the form of alphabet character representing the group of Tukey's HSD Post-Hoc calculation was automatically derived from statistics programs in LIMSMO system.

LIMSMO also provides comparative analysis between the published data as separate researches. Hence, we analyzed another locus in our LIMSMO database that belongs to transcription factor with non-transcription factor in separate research paper. Total 10 genes include 6 MoHOX genes were selected by comparison of *M. oryzae* ver.6 transcription factor data in fungal transcription factor database (Park *et al.*, 2008) (FTFD : http://ftfd.snu.ac.kr). We also selected a total of 10 non-transcription factor genes for comparison. Retrieved phenotype results of these 20 loci were expressed as an easily recognized form of list or table format. (Table 1).

After the experiment result input process, to suit our purpose, we grouped these genes to "Transcript Factor genes" and "Non-Transcription genes" at first, and we handled these results by using mask of searching result in "Public data" section. Making the filter mask for retrieving the result data that we want is very useful activities for research. After selecting the 20 loci, we made a filter mask set. For instance, if users want to find some mutant locus affecting mycelial growth level that is less than wildtype, activate and add "Growth Rate" search filter to filter masks, and click to select "Less" button at result filter option and include only non-chemically treated media (tCM, tMM, V8a, OMA, YEG) in "Growth Rate" search filter.As a result 9 mutant loci (MAGB; MGG_00365.6, MoSNF1; MGG_00803.6, MNH6; MGG_04489.6, MoHOX1; MGG_04853.6, MoCRZ1; MGG_05133.6, CHM1;

Table 1. Mutant phenotype defect in transcription factor and non-transcription factor genes

Locus Number	Gene Name	Strain	Phenotype Defect of Mutant[a]							
			GR[b]	PG	CN	GM	AP	CM[c]	PN[d]	PT
MGG_00184.6	MoHOX2	KJ201	99.5	-	0.0*	0.0*	0.0*	-	-	0.0*
MGG_04853.6	MoHOX1	KJ201	45.0*	-	83.7	100.3	100.9	Normal	-	93.3
MGG_06285.6	MoHOX4	KJ201	97.6	-	94.9	101.1	100.0	Small	-	80.0
MGG_11712.6	MoHOX6	KJ201	87.0*	-	105.7	102.0	100.3	Normal	-	66.7
MGG_12865.6	MoHOX7	KJ201	98.5	-	83.9	100.7	0.3*	Normal	-	0.0*
MGG_12958.6	MoHOX8	KJ201	93.8	-	90.2	102.6	98.8	Small	None	0.0*
MGG_02755.6	NUT1 (GATA)	Guy11	-	-	-	-	-	-	-	Normal
MGG_04489.6	MNH6 (HMG)	Guy11	Less	Less	56.4*	99.9	78.3*	Small	1 cell	Less
MGG_02696.6	MVP1 (C_2H_2)	70-15	-	Over	-	-	-	-	-	-
MGG_05133.6	MoCRZ1 (OB-fold)	KJ201	87.0*	-	67.3*	100.4	100.7	-	None	Less
MGG_00365.6	MAGB	70-15	80.0*	-	0.5*	-	2.6*	-	-	Less
MGG_00527.6	EMP1	70-15	Normal	-	104.0	99.0	55.5*	-	-	43.7*
MGG_00803.6	MoSNF1	70-15	76.1*	-	0.0*	79.1*	32.5*	Abnormal	-	Less
MGG_05201.6	MGB1	Guy11	89.1	-	11.2*	103.1	0.0*	-	None	Less
MGG_06320.6	CHM1	Guy11	54.5*	-	0.1*	Normal	Less	Abnormal	None	Less
MGG_06393.6	MgATG1	Guy11	Normal	-	6.6*	81.5*	94.5	-	None	Less
MGG_07456.6	MCNA	70-15	-	-	20.2*	97.4*	99.8*	-	-	56.3*
MGG_09100.6	CUT2	Guy11	Normal	-	Less	-	34.0*	-	-	Less

MGG_09898.6	MAC1	70-15	62.5*	-	14.4*	-	1.3*	-	-	Less
MGG_11899.6	MoCDC15	KJ201	78.8*	-	6.5*	19.2*	79.5*	-	None	Less

[a]Mutant phenotype is expressed as a percentage of what was observed in the wild type strain. Phenotypes significantly different (Tukey's HSD Post-Hoc calculation, significance level is 95%) from the wild type are noted by an asterisk. Abbreviation means GR:growth rate of fungal mycelium, PG:mycelial pigmentation, CN: ability to produce conidia, GM:conidial ermination, AP:appressorium formation, CM:conidial morphology, PN:appressorium penetration to host surface, PT:pathogenicity on host.

[b]Mycelial growth was measured only on non-chemically treated media.

[c]Small size and abnormal shape of conidial morphology are noted.

[d]Hypersensitive reaction observed in fungal penetration assay is expressed in "HR" term.

MGG_06320.6, MAC1;MGG_09898.6, MoHOX6; MGG_11712.6, MoCDC15; MGG_11899.6) is retrieved by search engine that applies above filter mask. Additionally, combination of two or more searching filter is also available. At first, we setup the primary filter mask that is the result of mutant mycelial growth level is normal and then selected the secondary filter mask that is appressorium formation level is less than wild type. In addition, we checked logical operation of filter mask set like this: growth rate of mutant is normal "And" appressorium formation of mutant is less than wild type. In this condition, the outcome was only 4 mutant loci (EMP1; MGG_00527.6, MGB1; MGG_05201.6, CUT2; MGG_09100.6, MoHOX7; MGG_12865.6).

The case study of the LIMSMO system shows the advantages of web-based system over a lab note in paper format. The direct connection between series of protocol steps and data of experimental results which are hard to express in paper form of lab note is the most valuable property of LIMSMO system. The second remarkable advantage of this system is easily recognizable genomic and genotype information of each locus that was performed because of system offers close linkage of phenotype data in LIMSMO system and genotype data in CFGP. The third advantage is flexibility and versatility of searching mask function that allows speedy searching of protocols used in experiment and selective finding at specific condition can be performed by just a few mouse clicks.

V. Gold standard for phenotype assays

We proposed the global standardized protocols for assaying *M. oryzae* phenotypes. The following protocols are stored in the LIMSMO protocol section, which can be accessed by other researchers.

1. Fungal strains and culture conditions

Wild-type strains and their transformants were grown on minimal agar medium (Talbot *et al.*, 1993), V8 agar (8% V8 juice (v/v), 1.5% agar (w/v), pH 6.79) or oatmeal agar (OMA; 5% oatmeal (w/v), 2% agar (w/v)) at 25°C under continuous fluorescent light.

2. Growth rate

Mycelial plugs (5×5 mm), excised from the colony edge of 4 day old culture on minimal agar, were put onto the center of culture plates. The media for measuring mycelial growth were complete agar and minimal agar as described previously (Talbot *et al.*, 1993). Cultures were incubated at room temperature (22-25°C) under constant white fluorescent light. Vegetative growth was measured by measuring the colony diameter at 4, 8 and 12 days post-inoculation when petri dishes (Φ90×h15 mm, SPL, cat. No. 10090) were used and at 2 and 4 days post-inoculation when 6-well cell culture plates (Φ35 mm, SPL, cat. No. 30006) were used. Each experiment involved hree replicates.

3. Mycelial pigmentation

For this experiment, mycelial plugs (see section of growth rate check assay) were placed on the center of V8 juice agar. The bottom side of culture plate was photographed at 12 days post-inoculation for cultures grown on petri plate and at 7 days post-inoculation for those grown in 6 well cell culture plates.

4. Conidiation

Conidia production was measured by counting the number of conidia from 12-days-old (90 mm petri dish) or 7-days-old (6 well cell culture plate) V8 juice agar cultures. For harvesting conidia, sterile distilled water was poured into the plate surface and scraped with glass rod or 1.5 ml microcentrifuge-tube tip (for 90 mm petri dish, use 20ml of ddH_2O and for 6 well cell culture plate, pour 5ml of ddH_2O). The number of conidia in 10 µl of conidial suspension was counted using a hemacytometer with three replicates. Counting is repeated three times.

5. Conidial germination and appressorium formation

Conidial germination and appressorium formation were measured on microscope cover slip or hydrophilic and hydrophobic side of GelBond® film (FMC Corporation). Conidia harvested for quantifying conidial production were filtered through Miracloth®. A conidial suspension of 40 µl was pipetted onto an assay surface following the adjustment of its concentration to approximately 5×10^4 conidia/mL. Drops were placed in a moistened box and incubated at 25°C. The percentage of

conidia germinating and germinated conidia-forming appressoria was determined by microscopic examination of at least 100 conidia per replicate in at least three independent experiments, with three replicates per experiment. Determination conidia germination was counted at 2 and 4 hours after post-inoculation and conidia forming appressoria was counted at 8 and 16 hours after post-inoculation.

6. Conidial morphology

Conidial size was measured under a microscope. Abnormal morphology of conidia compared with wild-type was also recorded.

7. Pathogenicity

For the pathogenicity assay, conidia were harvested from 7 (in 6-well cell culture plate) to 12 (in 90mm petri dish) -day-old cultures on oatmeal agar or V8 juice agar, and 10 mL of conidial suspension (250 ppm Tween 20; 1×10^5 spores/ml) filtered through Miracloth® were sprayed onto susceptible rice seedlings (*Oryza sativa* cv. Nakdongbyeo; *Oryza sativa* cv. Lijiangxintuanheigu (LTH); *O. sativa* cv. CO-39; *O. sativa* cv. S-201). The inoculated rice seedlings were kept in a dew chamber at 25°C under dark and with 100% humidity for 24 h, and then transferred to a greenhouse or incubator with a photoperiod of 16 hours with fluorescent light. Disease severity (Table S3 in supplement data) (IRRI, 1996) was measured at 7 days after inoculation, as previously described (Valent *et al.*, 1991). For the infiltration infection assay, 100 μl of conidia suspension was inoculated into wounded leaves at three points per leaf

of 4-week-old rice plants using syringe without needle. These experiments were replicated three times.

8. Appressorium penetration

This assay was performed using onion epidermis and rice sheaths. For preparing onion epidermis, the second outer layers of yellow onions were prepared and rinsed with water. The onion bulb scale was not washed with chloroform. The inner epidermis was peeled off, cut into 1 cm^2 strips and floated on distilled water. About 400 conidia in 20 µl were placed on the strips (Xu *et al.*, 1997). For the sheaths infection assay, leaf sheaths of the third and fourth leaves of rice plants at the 4.5 leaf stage were peeled off, and a length from the middle part of each leaf sheath was removed. The excised leaf sheath pieces were filled with a conidial suspension (1-5 ml, 2×10^4 conidia/ml) of *M. oryzae* using a pipet tip (Koga *et al.*, 2004). The samples were incubated at 25°C in transparent boxes in humid condition under white fluorescent light with constant light period. After incubation, the penetrated samples were fixed with lactophenol (lactic acid/phenol/glycerol/water = 1:1:1:1) and infection hyphae and host cell auto fluorescence were observed using differential interference contrast microscope. Appressorium formation and infectious hyphae growth were examined after 24, 48 and 72 hours of incubation at room temperature.

CONCLUSION AND DISCUSSION

As the enormous amount of genome data spurting out and functional genomics becomes more advanced, large-scale phenomics encountered challenges in a new way for the efficient data handling. For this reason, laboratory information management system has been implemented to support managing and querying the phenotype screening data for several model organisms. For example, qualitative and quantitative phenotype data from the model plant *Arabidopsis thaliana* (Lu *et al.*, 2011) and standardized phenotype information from large scale insertional mutagenesis project of rice are managed by LIMS software (Henry *et al.*, 2008). For the model microbial organisms, LIMS for mutant phenotype information is developed in a part of *Saccharomyces* Genome Database project (Costanzo *et al.*, 2009) and PACLIMS platform tracking and storing both biological materials and result data is constructed for *M. oryzae* (Donofrio *et al.*, 2005).

As stated above, there were several challenges in making LIMS platform of biological assay for model organisms. These systems have the same disciplines for curating and managing information of targeted organism and phenotype assay data. But, these commodities are different among the platform because each organism has different guidelines for getting information from phenotype assay. On that point, we have developed a new leading LIMS platform for *M. oryzae* to manage various phenotype/genotype data.

LIMS for *M. oryzae* has been indispensable resources enabling one to overview and share the experimental data of characterized genes and have a glimpse of protocol and methods used in phenotypic assay. It is an essential tool for researchers who study rice blast fungus. This system provides guidelines for phenotype assay in *M. oryzae* and follows experimental workflow and formatted result in form of experimental assay and suggests a widely accepted standard protocol derived from the most used methods in previously published papers. First, in project setup phase, related papers collected in literature research are stored and managed by our publication management component. In the generation of mutant phase, genotype information of selected and being researched one is managed by mutant management component that is linked to CFGP genome data warehouse. Next, researchers would get formatted result data of mutant phenotype experiment by using the global standardized protocol and method which is defined and stored in LIMSMO protocol component. These experimental data is stored by predefined format in LIMSMO database and analyzed by external statistic program automatically. In the final step, all of the data can be searched and shared by all peers or community members.

Although many previously published experimental results data in the system have heterogeneous materials and methods, LIMSMO system provides grouping function for experimentally related locus and searching mask function for flexible and speedy searching of specific experimental condition and protocols of method. Moreover, the system implemented with a new standardized guideline of phenotype assays and data acquisition formats for peers in the same research field.

To advance our knowledge of fungal pathogenomics, LIMSMO has three significant improved developments. First, researchers can input new data retrieved from assay through the web in LIMSMO platform, while previously developed systems of microbe LIMS platform curate only existing data from experimental result. Second, researchers follow standardized guideline of data input and statistic information for these data is automatically calculated. With the comparable information, peers can compare the stored experimental data by using quantitative analysis. Third improvement of this platform is the close linkage of phenotype and genotype data. All phenotype information stored our system is connected to CFGP by unique id that has genome sequence information, so users can carry out further bioinformatics analysis by using various tools in CFGP (Choi *et al.*, 2013).

To conclude, LIMSMO is developed for integrating the experimental data and electronic database to provide the environment of sharing and comparing the data processed by standardized protocol between the peers. This system is a valuable resource to serve the community who work on rice blast disease. Furthermore, it will shed light on a model platform for pathogenomics, not only *M. oryzae* but also plant pathogenic fungi, to understand the disease mechanism at the systems level.

LITERATURE CITED

Balhadere, P.V., Foster, A.J., Talbot, N.J. (1999) Identification of pathogenicity mutants of the rice blast fungus *Magnaporthe grisea* by insertional mutagenesis. *Mol Plant Microbe In* **12**: 129-142.

Chen, X.L., Yang, J., Peng, Y.L. (2011) Large-scale insertional mutagenesis in *Magnaporthe oryzae* by Agrobacterium tumefaciens-mediated transformation. *Methods in molecular biology* **722**: 213-224.

Chi, M.H., Park, S.Y., Kim, S., Lee, Y.H. (2009) A Novel Pathogenicity Gene Is Required in the Rice Blast Fungus to Suppress the Basal Defenses of the Host. *Plos Pathog* **5**.

Choi, J., Cheong, K., Jung, K., Jeon, J., Lee, G.W., Kang, S., Kim, S., Lee, Y.W., Lee, Y.H. (2013) CFGP 2.0: a versatile web-based platform for supporting comparative and evolutionary genomics of fungi and Oomycetes. *Nucleic acids research* **41**: D714-D719.

Costanzo, M.C., Skrzypek, M.S., Nash, R., Wong, E., Binkley, G., Engel, S.R., Hitz, B., Hong, E.L., Cherry, J.M., the Saccharomyces Genome Database, P. (2009) New mutant phenotype data curation system in the *Saccharomyces* Genome Database. *Database : the journal of biological databases and curation* **2009**: bap001.

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.R., Pan, H., Read, N.D., Lee, Y.H., Carbone, I.,

Brown, D., Oh, Y.Y., Donofrio, N., Jeong, J.S., Soanes, D.M., Djonovic, S., Kolomiets, E., Rehmeyer, C., Li, W., Harding, M., Kim, S., Lebrun, M.H., Bohnert, H., Coughlan, S., Butler, J., Calvo, S., Ma, L.J., Nicol, R., Purcell, S., Nusbaum, C., Galagan, J.E., Birren, B.W. (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434**: 980-986.

Donofrio, N., Rajagopalon, R., Brown, D., Diener, S., Windham, D., Nolin, S., Floyd, A., Mitchell, T., Galadima, N., Tucker, S., Orbach, M.J., Patel, G., Farman, M., Pampanwar, V., Soderlund, C., Lee, Y.H., Dean, R.A. (2005) 'PACLIMS': a component LIM system for high-throughput functional genomic analysis. *BMC Bioinformatics* **6**: 94.

Droit, A., Hunter, J.M., Rouleau, M., Ethier, C., Picard-Cloutier, A., Bourgais, D., Poirier, G.G. (2007) PARPs database: a LIMS systems for protein-protein interaction data mining or laboratory information management system. *BMC Bioinformatics* **8**: 483.

Ebbole, D.J. (2007) *Magnaporthe* as a model for understanding host-pathogen interactions. *Annual review of phytopathology* **45**: 437-456.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G. (1996) Life with 6000 genes. *Science* **274**: 546, 563-547.

Henry, L., Ramm, K., Zhu, Q.H., Upadhyaya, N.M. (2008) RGMIMS: a web-based laboratory information management system for plant functional genomics

research. *Mol Breeding* **22**: 151-157.

IRRI (1996) Standard Evaluation System for rice 4th edition. *IRRI*: 17.

Jeon, J., Park, S.Y., Chi, M.H., Choi, J., Park, J., Rho, H.S., Kim, S., Goh, J., Yoo, S., Choi, J., Park, J.Y., Yi, M., Yang, S., Kwon, M.J., Han, S.S., Kim, B.R., Khang, C.H., Park, B., Lim, S.E., Jung, K., Kong, S., Karunakaran, M., Oh, H.S., Kim, H., Kim, S., Park, J., Kang, S., Choi, W.B., Kang, S., Lee, Y.H. (2007) Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. *Nature genetics* **39**: 561-565.

Kim, S., Park, S.Y., Kim, K.S., Rho, H.S., Chi, M.H., Choi, J., Park, J., Kong, S., Park, J., Goh, J., Lee, Y.H. (2009) Homeobox transcription factors are required for conidiation and appressorium development in the rice blast fungus *Magnaporthe oryzae*. *Plos Genet* **5**: e1000757.

Koga, H., Dohi, K., Nakayachi, O., Mori, M. (2004) A novel inoculation method of *Magnaporthe grisea* for cytological observation of the infection process using intact leaf sheaths of rice plants. *Physiol Mol Plant P* **64**: 67-72.

Kour, A., Greer, K., Valent, B., Orbach, M.J., Soderlund, C. (2012) MGOS: development of a community annotation database for *Magnaporthe oryzae*. *Molecular plant-microbe interactions : MPMI* **25**: 271-278.

Lu, Y., Savage, L.J., Larson, M.D., Wilkerson, C.G., Last, R.L. (2011) Chloroplast 2010: a database for large-scale phenotypic screening of Arabidopsis mutants. *Plant physiology* **155**: 1589-1600.

Park, J., Park, J., Jang, S., Kim, S., Kong, S., Choi, J., Ahn, K., Kim, J., Lee, S., Kim,

S., Park, B., Jung, K., Kim, S., Kang, S., Lee, Y.H. (2008) FTFD: an informatics pipeline supporting phylogenomic analysis of fungal transcription factors. *Bioinformatics* **24**: 1024-1025.

Robinson, A.L. (1983) LIMS Is next step in laboratory automation. *Science* **220**: 180-183.

Saal, L.H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A., Peterson, C. (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome biology* **3**: SOFTWARE0003.

Sanchez-Villeda, H., Schroeder, S., Polacco, M., McMullen, M., Havermann, S., Davis, G., Vroh-Bi, I., Cone, K., Sharopova, N., Yim, Y., Schultz, L., Duru, N., Musket, T., Houchins, K., Fang, Z., Gardiner, J., Coe, E. (2003) Development of an integrated laboratory information management system for the maize mapping project. *Bioinformatics* **19**: 2022-2030.

Skobelev, D.O., Zaytseva, T.M., Kozlov, A.D., Perepelitsa, V.L., Makarova, A.S. (2011) Laboratory information management systems in the work of the analytic laboratory. *Meas Tech+* **53**: 1182-1189.

Soderlund, C., Haller, K., Pampanwar, V., Ebbole, D., Farman, M., Orbach, M.J., Wang, G.L., Wing, R., Xu, J.R., Brown, D., Mitchell, T., Dean, R. (2006) MGOS: A resource for studying *Magnaporthe grisea* and *Oryza sativa* interactions. *Molecular plant-microbe interactions : MPMI* **19**: 1055-1061.

Talbot, N.J. (2003) On the trail of a cereal killer: Exploring the biology of

Magnaporthe grisea. *Annu Rev Microbiol* **57**: 177-202.

Talbot, N.J., Ebbole, D.J., Hamer, J.E. (1993) Identification and characterization of MPG1, a gene involved in pathogenicity from the rice blast fungus *Magnaporthe grisea*. *Plant Cell* **5**: 1575-1590.

Valent, B., Farrall, L., Chumley, F.G. (1991) *Magnaporthe grisea* genes for pathogenicity and virulence identified through a series of backcrosses. *Genetics* **127**: 87-101.

Voegele, C., Tavtigian, S.V., de Silva, D., Cuber, S., Thomas, A., Le Calvez-Kelm, F. (2007) A laboratory information management system (LIMS) for a high throughput genetic platform aimed at candidate gene mutation screening. *Bioinformatics* **23**: 2504-2506.

Xu, J.R., Urban, M., Sweigard, J.A., Hamer, J.E. (1997) The CPKA gene of *Magnaporthe grisea* is essential for appressorial penetration. *Mol Plant Microbe In* **10**: 187-194.

Supplement Table 1. Locus tagged gene and task of eight phenotype assay information in each paper

Author (Year)	Accession Number[a]	Locus Number	Gene Name	Number of Tasks in Phenotype Assay[b]							
				GR	PG	CN	GM	AP	CM	PN	PT
Adachi, K. 1998	9707535	MGG_09898.6*	MAC1	-	-	-	-	4	-	-	1
Ahn, N. 2004	15055545	MGG_00527.6	EMP1	1	-	1	1	5	-	-	1
Asiegbu, F. O. 2004	14734169	MGG_02696.6	MVP1	-	1	-	-	-	-	-	-
Avila-Adame, C. 2002	12036280	MGG_12936.6	AOXMg	1	-	-	-	-	-	-	-
Balhadere, P. V. 2001	11549759	MGG_00111.6	PDE1	-	-	-	-	-	-	-	2
Bhambra, G. K. 2006	16824094	MGG_01721.6	PTH2	4	-	-	-	-	-	1	2
Bruno, K. S. 2004	15590826	MGG_09565.6*	PMK1	-	-	-	-	1	-	-	1
Chi, M. 2009	19390617	MGG_04163.6	DES1	1	1	1	1	1	1	1	1
Choi, J. 2009	19111943	MGG_05133.6	MoCRZ1	5	-	1	1	1	-	2	1
Choi, J. 2009	19190403	MGG_07456.6	MCNA	4	-	1	5	5	-	-	1
Choi, J. 2011	21237279	MGG_05332.6	MoPLC2	2	-	1	-	-	-	-	3
		MGG_08315.6	MoPLC3	2	-	1	-	-	-	-	3
Choi, W. 1997	9401122	MGG_09898.6*	MAC1	1	-	2	-	8	1	-	2
Clergeot, P. H. 2001	11391010	MGG_12594.6	PLS1	-	-	-	-	-	-	1	1
DeZwaan, T. M. 1999	10521529	MGG_05871.6	PTH11	-	-	-	-	6	-	1	3
Dixon, K. P. 1999	10521531	MGG_01822.6	OSM1	2	-	-	-	-	-	-	1
Egan, M. J. 2007	17600089	MGG_00750.6	NOX1	1	-	-	-	-	-	1	1

		MGG_06559.6	NOX2	1	-	-	-	-	-	-	1
Foster, A. J. 2003	12514128	MGG_03860.6*	TPS1	16	-	-	-	-	-	-	1
		MGG_09471.6	NTH1	4	-	-	-	-	-	-	1
		MGG_01261.6	TRE1	5	-	-	-	-	-	-	1
Froeliger, E. H. 1996	8757395	MGG_02755.6	NUT1	1	-	-	-	-	-	-	1
Goh, J. 2011	21600998	MGG_11899.6	MoCDC15	4	-	1	2	1	-	1	1
Gupta, A. 2007	17462923	MGG_16255T0	MGA1	2	-	-	1	1	-	-	3
Gupta, A. 2008	18034832	MGG_00937.6	ABC4	-	-	-	-	1	-	1	2
Jeon, J. 2008	18393612	MGG_00883.6	MCK1	5	-	1	1	1	-	1	2
Jeong, J. S. 2007	17590228	MGG_07075.6	MSP1	-	-	-	-	-	-	-	2
Kamakura, T. 2002	12036274	MGG_12939.6	CBP1	-	-	-	-	3	-	-	2
Kim, S. 2005	16101997	MGG_01173.6	MHP1	3	-	3	1	1	1	1	1
Kim, S. 2009	19997500	MGG_04853.6	MoHOX1	1	-	1	1	1	1	-	1
		MGG_00184.6	MoHOX2	1	-	1	1	1	1	-	1
		MGG_01730.6	MoHOX3	1	-	1	1	1	1	-	1
		MGG_06285.6	MoHOX4	1	-	1	1	1	1	-	1
		MGG_07437.6	MoHOX5	1	-	1	1	1	1	-	1
		MGG_11712.6	MoHOX6	1	-	1	1	1	1	-	1
		MGG_12865.6	MoHOX7	1	-	1	1	1	1	-	1
		MGG_12958.6*	MoHOX8	1	-	1	1	1	1	-	1
Kwon, M. 2010	ISI:000275103100002	MGG_00056.6	MoSDR1	1	-	1	1	1	-	1	1

Lau, G. W. 1998	9742203	MGG_09847.6	ACR1	-	-	-	-	1	1	-	1
Li, L. 2004	15141959	MGG_06320.6	CHM1	2	-	1	1	1	1	1	2
		MGG_12821.6	MST20	2	-	1	-	-	-	1	1
Li, L. 2007	17427815	MGG_02370.6	MIR1	1	-	1	2	2	-	2	1
Liu, H. 2007	17255942	MGG_14517.6	RGS1	-	-	-	-	2	-	-	-
Liu, S. 1997	9390422	MGG_01818.6	MAGA	1	-	1	-	1	-	-	-
		MGG_00365.6	MAGB	3	-	1	-	1	-	-	1
		MGG_04204.6	MAGC	1	-	1	-	1	-	-	-
Liu, X. H. 2007	17416896	MGG_06393.6	MgATG1	2	-	1	1	1	-	1	2
Lo, S. C. 2002	12455965	MGG_07384.6	CBS1	1	-	-	-	-	-	-	-
Lu, J. P. 2007	17644013	MGG_04489.6	MNH6	1	1	1	1	1	1	1	1
Mitchell, T. K. 1995	8535140	MGG_06368.6	CPKA	-	-	-	-	4	-	-	1
Mori, T. 2008	18218020	MGG_08938.6	MDG1	2	-	-	2	-	-	-	1
Motoyama, T. 2005	15707841	MGG_11174.6	HIK1	8	-	-	-	1	-	-	1
Nishimura, M. 2003	14507377	MGG_05201.6	MGB1	1	-	1	1	2	-	1	2
Park, G. 2002	11952120	MGG_12958.6*	MST12	1	-	1	1	1	1	1	2
Park, G. 2004	15341648	MGG_12958.6*	MST12	-	-	-	-	-	-	1	1
Rho, H. S. 2009	19400837	MGG_02444.6	MoPLC1	1	-	1	1	1	-	1	2
Shi, Z. 1998	9487695	MGG_05287.6	CON7	1	-	-	-	2	1	-	1
Skamnioti, P. 2007	17704215	MGG_09100.6	CUT2	4	-	1	-	9	-	-	2
Sweigard, J. A. 1992	1557024	MGG_01943.6	CUT1	-	-	-	-	-	-	-	1

Talbot, N. J. 1993	8312740	MGG_10315.6	MPG1	-	-	-	-	1	-	-	1
Viaud, M. C. 2002	11971145	MGG_10447.6	CYP1	2	-	-	-	2	-	-	1
Villalba, F. 2008	17716934	MGG_10157.6	MgKU80	2	-	1	1	-	-	-	1
Wang, Z. Y. 2003	12622815	MGG_04895.6	ICL1	3	-	1	1	1	-	1	2
Wilson, R. A. 2007	17641690	MGG_03860.6*	TPS1	8	-	1	-	-	-	-	1
Wu, S. C. 2006	16461639	MGG_02245.6	XYL6	1	-	-	-	-	-	-	1
Xu, J. R. 1996	8946911	MGG_09565.6*	PMK1	-	-	1	1	1	-	-	1
Yi, M. 2008	ISI:000256316000005	MGG_08180.6	MHF21	-	-	1	1	1	-	1	1
		MGG_06766.6	MHF16	-	-	1	1	1	-	1	1
Yi, M. 2008	18595748	MGG_00803.6	MoSNF1	13	-	1	1	1	1	-	2
Yi, M. 2009	19252083	MGG_06648.6	LHS1	3	2	1	2	2	-	2	2
Zhao, X. 2007	17214742	MGG_00800.6	MST7	-	-	-	-	2	-	-	-
Zheng, W. 2007	17933908	MGG_10323.6	MgRHO3	1	-	1	-	1	1	1	1
Total 57 Papers		Total 67(72) Locus[c]		139	5	43	40	92	18	28	89

[a]Exclude two papers that tagged ISI web of knowledge index, all accession number is delivered from PubMed library(PMID).

[b]Abbreviation below phenotype assay means GR:Growth Rate, PG:Pigmentation, CN:Conidiation, GM:Germination, AP:Appressorium Formation, CM:Conidial Morphology, PN:Penetration, PT:Pathogenicity.

[c,*]We collect total 72 genes in paper of mutant phenotype assay. But 4 genes were researched again by another paper.

Supplement Table 2-1. Number of used material and method for assay of appressorium formation check

Protocol type	Material & method	# of used
Measurement methods	Count more than 100 conidia	70
	Count more than 120 conidia	9
	Count more than 80 conidia	4
	Count more than 150 conidia	1
	Count more than 300 conidia	1
	Not specified	7
Surface type	Plastic microscope cover slip	40
	GelBond® film (hydrophobic side)	26
	GelBond® film (hydrophilic side)	15
	Barley cultivar Bonanza	3
	Polycarbonate surface	3
	Hydrophilic glass slides	3
	Teflon surfaces	2
Conidial suspensions	40 µl of 5×10^4 conidia/ml	36
	50 µl of 1×10^5 conidia/ml	19
	1×10^4 conidia/ml	17
	2×10^4 conidia/ml	10
	50 µl of 2×10^5 conidia/ml	4
	50 µl of 3×10^4 conidia/ml	1
	1×10^3 conidia/ml	1
	Not specified	4
Observation times	24 hour post-inoculation	23
	16 hour post-inoculation	21
	9 hour post-inoculation	12
	12, 24, 48 hour post-inoculation	9
	20 hour post-inoculation	5
	8 hour post-incubation	5
	12 hours post-inoculation	4
	0.5, 2, 12, 24 hour post-inoculation	2
	12, 24, 36 hour post-inoculation	1
	2, 4, 6, 12, 24, 48 hour post-inoculation	1

	18 hour post-inoculation	1
	6, 8, 12, 24 hour post-inoculation	1
	2, 4, 8, 24 hour post-inoculation	1
	4, 8, 24, 48 hour post-inoculation	1
	2, 10, 20 hour post-inoculation	1
	2 hour post-inoculation	1
	4, 6, 8 hour post-inoculation	1
	8, 16, 24 hour post-inoculation	0
	6 hour post-inoculation	0
	2, 4, 8, 12, 24 Hour post-inoculation	0
	Not specified	2
Treatment methods	No treatment	56
	cAMP 10 mM	7
	IBMX 2.5 mM	7
	DAG 20 ppm	5
	Diol 100 µM	4
	monobutytyl-cAMP 5 mM	2
	Diol 1 µM	2
	Diol 10 µM	1
	cAMP 1 mM	1
	Propranolol 1 µM	1
	CsA	1
	cAMP 100 µM	1
	$CaCl_2$ 100 mM	1
	Cold Shock (90 min, 4°C)	1
	cAMP 50 mM	1

Supplement Table 2-2. Number of used material and method for assay of sporulation/conidiation check

Protocol type	Material & method	# of used
Collection methods	Scrape harvesting	39
	Not specified	4
Measurement methods	Count by microscope	39
	Not specified	4
Media type	Oatmeal agar medium	35
	Complete medium	3
	V8 juice agar medium	2
	Rice polish agar medium	1
	Minimal medium + plant cutin	1
	Glucose minimal medium	1
	Not specified	0
Plate type	Petri dish (90 mm)	31
	6-well plate (36 mm)	4
	Mycelial disc, 1 cm diameter	1
	Petri dish (150 mm)	1
	Not specified	6
Observation times	10 days post-inoculation	13
	12 days post-inoculation	11
	5 days post-inoculation	5
	8 days post-inoculation	3
	7 days post-inoculation	3
	14 days post-inoculation	2
	11 days post-inoculation	2
	16 days post-inoculation	1
	6 days post-inoculation	0
	Not specified	3

Supplement Table 2-3. Number of used material and method for assay of conidial germination check

Protocol type	Material & method	# of used
Measurement method	Count more than 100 conidia	33
	Count more than 200 conidia	2
	Count more than 150 conidia	1
	Count more than 300 conidia	1
	Not specified	3
Surface type	Plastic microscope cover slip	25
	GelBond® film (hydrophobic side)	13
	GelBond® film (hydrophilic side)	1
	3% agar plates	1
Conidial suspension	40 µl of 5×10^4 conidia/ml	24
	2×10^4 conidia/ml	6
	40 µl of 1×10^4 conidia/ml	2
	20 µl of 1×10^5 conidia/ml	2
	3×10^4 conidia/ml	2
	1×10^6 conidia/ml	1
	1×10^3 conidia/ml	1
	Not specified	2
Observation times	16 hour post-inoculation	12
	9 hour post-inoculation	8
	12 hours post-inoculation	4
	8 hours post-incubation	3
	2 hours post-inoculation	3
	3, 6, 9, 12, 18 hour post-inoculation	2
	2, 4 hour post-inoculation	2
	0.5, 2, 12, 24 hour post-inoculation	2
	2, 4, 8, 12, 24 hour post-inoculation	1
	2, 4, 8, 24 hour post-inoculation	1
	2, 10, 20 hour post-inoculation	1
	2, 4, 6, 12, 24, 48 hour post-inoculation	1
Treatment methods	No treatment	34

cAMP 100 µM	1
Diol 100 µM	1
DAG 20 ppm	1
CaCl2 100 mM	1
Cold Shock (90 min, 4°C)	1
50% CM	1

Supplement Table 2-4. Number of used material and method for assay of mycelial growth rate check

Protocol type	Material & method	# of used
Inoculum source	Point inoculation	78
	Mycelial block (5 mm in diameter from MM)	36
	Mycelial disc (7 mm in diameter)	5
	Mycelial block (5 mm in diameter from V8a)	4
	Mycelial agar plug (4 mm in diameter)	3
	Mycelial block (5 mm in diameter from CM)	2
	$3{\times}10^3$ conidia	2
	Mycelial plugs (2 mm in diameter)	1
	Not specified	8
Light condition	White fluorescent light	116
	Not specified	23
Light cycle	Constant light	110
	12 hour Dark / 12 hour Light Cycle	6
	Not specified	23
Media type	Modified CM for *M. oryzae* (Talbot et al.)	37
	Oatmeal agar medium	10
	MM + sole carbon source glucose	4
	Modified V8 juice agar medium	4
	MM + sodium acetate	4
	Modified MM for *M. oryzae* (Talbot et al.)	3
	MM + olive oil	3
	MM + mannitol	3
	Potato dextrose agar medium	3
	MM + sole carbon source fructose	2
	MM + sole carbon source sucrose	2
	MM + sole carbon source maltose	2
	CM + calcofluor white 100 ppm	2
	CM + calcofluor white	2
	MM + sole carbon source trehalose	2
	CM + SDS 0.01 M	1
	MM + sole carbon source Xylose	1

MM + sole carbon source Arabinose	1
MM + sole carbon source Galactose	1
MM + sole carbon source Xylan	1
MM + sole carbon source Pectin	1
MM + sole carbon source RCW	1
MM + sole carbon source RCW + Glu	1
Minimal medium (pH 4.5)	1
Minimal medium (pH 6.5)	1
Minimal medium (pH 8.5)	1
Minimal medium (pH 10.5)	1
CM + copper sulfate 4 mM	1
CM + manganase chloride 10 mM	1
MM - carbon-starved medium	1
MM - nitrogen-starved medium	1
MM + hydrogen peroxide	1
MM + lysing enzyme	1
MM + hydrogen peroxide + lysing Enzyme	1
CM + azoxystrobin	1
MM + Glucose 50 mM	1
MM + sodium acetate 50 mM	1
MM + triolein	1
CM + NaCl 0.4 mM	1
MM + sodium acetate + glucose	1
MM + mannitol + glucose	1
MM + glycerol + glucose	1
MM + ethanol + glucose	1
MM + casamino acid + sodium acetate	1
MM + vitamin solution + glucose	1
MM + different concentrations of homocysteine(HCY)	1
YG agar mediaum	1
YGA + sorbitol 1 M	1
YGA + glycerol 1 M	1
YGA + NaCl 0.5 M	1
YGA + KCl 0.5 M	1
YGA + fludioxonil 0.2 ppm	1

	YGA + iprodione 12.5 ppm	1
	YGA + PCNB 12.5 ppm	1
	YEG agar medium	1
	MM - glucose	1
	MM + plant cutin	1
	CM with CsA	1
	Fructose MM with NO_3^-	1
	Fructose MM with NH_4^+	1
	Glucose MM with NO_3^-	1
	CM + 0.2M $CaCl_2$	1
	Glucose MM with NH_4^+	1
	Glucose MM with NO_2^-	1
	Glucose MM with cysteine	1
	Glucose MM with $KClO_3^-$ + NH_4^+	1
	Glucose MM with hypoxanthine	1
	MM - nitrogen-starved medium + various nitrogen source 10 mM	1
	CM + sorbitol	1
	CM + NaCl 0.5 M	1
	CM + $MnCl_2$ 0.02 M	1
	N-starvation medium (CM-N source)	0
	C-starvation medium (CM-C source)	0
Plate type	Petri dish (90 mm)	101
	6-well plate (36 mm)Petri dish (90 mm)	10
	Race-tube	5
	Liquid media tube	1
	Not specified	22
Growth temperature	RT	117
	26°C	11
	24°C	3
	28°C	2
	Not specified	6
Observation times	12 days post-inoculation	43
	10 days post-inoculation	30
	2.5-10 days post-inoculation	8
	4, 8, 12 days post-inocluation	8

9 days post-inoculation	8
5 days post-inoculation	7
6 days post-inoculation	5
6, 12 days post-inoculation	5
Average growth per day	4
4 days post-inoculation	3
5, 7, 10 days post-inoculation	3
7 days post-inoculation	3
8 days post-inoculation	2
3 days post-inoculation	2
11 days post-inoculation	1
3, 5, 7, 9 days post-inoculation	1
21 days post-inoculation	1
Not specified	5

Supplement Table 2-5. Number of used material and method for assay of conidial morphology check

Protocol type	Material & method	# of used
Measurement method	Microscope	18

Supplement Table 2-6. Number of used material and method for assay of pathogenicity check

Protocol type	Material & method	# of used
Infection host	Rice cultivar Nakdong (4th leaf stage)	31
	Rice cultivar CO-39 (4th leaf stage)	30
	Barley cultivar Golden Promise	8
	Rice cultivar S-201	4
	Barley cultivar Bonanza	3
	Rice cultivars Maratelli and Sariceltic	2
	Rice cultivar Koshihikari	2
	Barley leaf segments	2
	Rice (unknown cultivar)	1
	Rice cultivar Koshihikari	1
	Rice cultivar Nihonbare	1
	Rice cultivar spl11	1
	Rice cultivar M103	1
	Rice cultivar Yashiro-mochi	1
	Rice cultivars Bala, C101Lac	1
Inoculation method	Spray inoculation	75
	Infiltration method	9
	Injection inoculation	4
Inoculum type	10 ml of 1×10^5 conidia/ml	49
	5×10^5 spores/ml	9
	1 ml of 5×10^4 conidia/ml	6
	10 ml of 2×10^5 conidia/ml	6
	100~200 µl of $2\sim3\times10^4$ conidia/ml	5
	1×10^4 conidia/ml	5
	1×10^6 conidia/ml	4
	3×10^5 conidia/ml	2
	10 ml of 3×10^4 conidia/ml	1
	Not specified	2

Light condition	1 day dark (humid condition), 16 hours light/8 hours dark cycle	79
	14 hours light / 10 hours dark	1
	Not specified	9
Measurement method	Qualitative description	69
	Disease index	15
	Lesion ratio	2
	The percentages of diseased leaf area (DLA) and the proportion of DLA based on lesion types	2
	Counting lesion number	1
Observation time	7 days post-inoculation	63
	5 days post-inoculation	9
	10 days post-inoculation	5
	8 days post-inoculation	4
	6 days post-inoculation	2
	4~5 days post-inoculation	2
	4~6 days post-inoculation	1
	5~6 days post-inoculation	1
	Not specified	2
Treatment method	No treatment	83
	cAMP 10 mM	3
	Osmotic stabilizer (KCl or NaCl)	2
	Diacylglycerol (DAG) 1,2-diocatanoylglycerol-rac-glycerol	1

Supplement Table 2-7. Number of used material and method for assay of penetration check

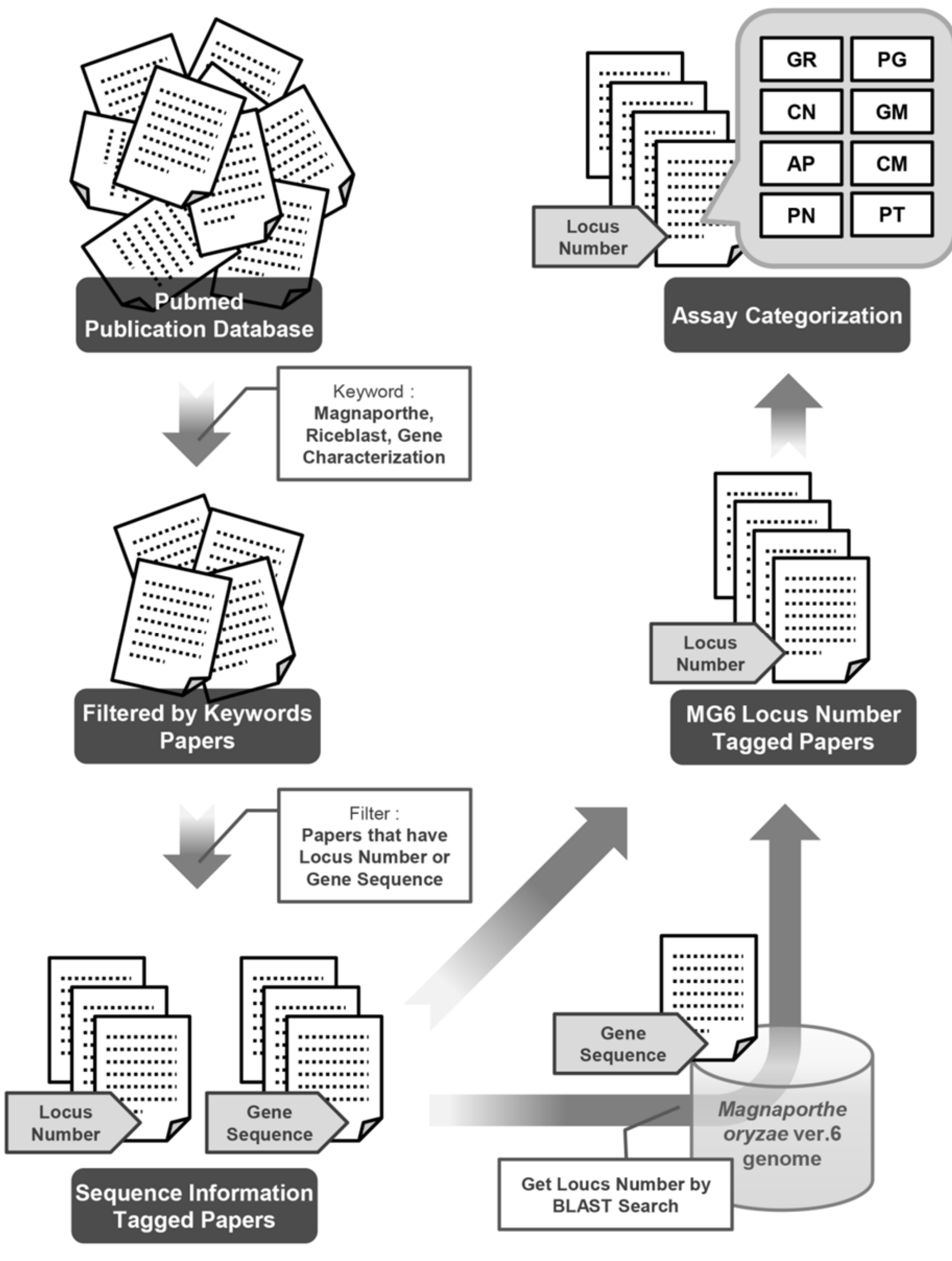
Protocol type	Material & method	# of used
Measurement method	Microscope	28
Surface type	Onion epidermis	17
	Rice sheath	6
	Barley leaf (cultivar BarSoy)	3
	Rice leaf	1
	PUDO-193 cellophane	1
Observation time	48 hours post-inoculation	13
	72 hours post-inoculation	3
	48~72 hours post-inoculation	3
	96 hours post-inoculation	2
	24 hours post-inoculation	2
	24, 48, 72, 96 hours post-inoculation	2
	48~96 hours post-inoculation	1
	120 hours post-inoculation	1
	112 hours post-inoculation	1

Supplement Table 2-8. Number of used material and method for assay of pigmentation check

Protocol type	Material & method	# of used
Inoculum source	Mycelial block (6 mm in diameter from MM)	2
	Point inoculation	1
	Mycelial disc (7 mm in diameter)	1
	Mycelium disc (5 mm in diameter)	1
Media type	Modified CM for *M. oryzae* (Talbot et al.)	2
	CM + $MnCl_2$ 10 mM	1
	Oatmeal agar medium	1
	PDA + ferulic acid 0.01%	1
Plate type	Petri dish (90 mm)	5

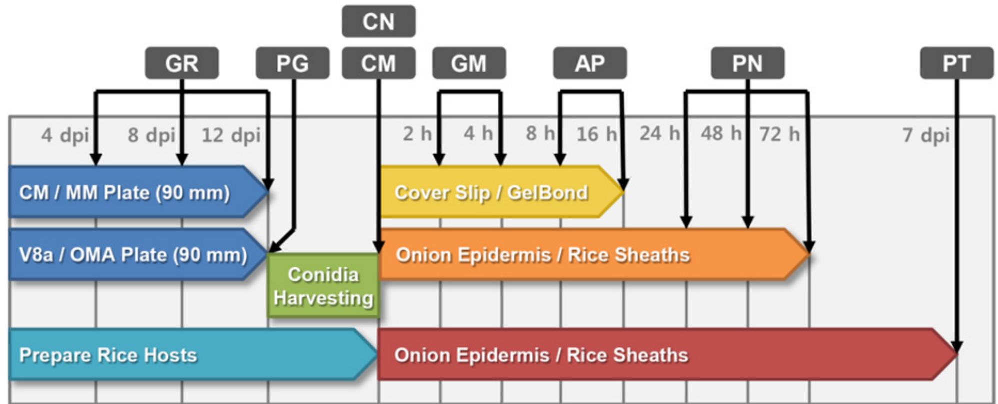
Supplement Table 3: Disease severity standard evaluation system of IRRI (1996)

Disease score	**Lesion area (%)**	**Disease reaction**
1	1-5	R
3	6-12	MR
5	13-25	MS
7	26-50	S
9	51-100	HS



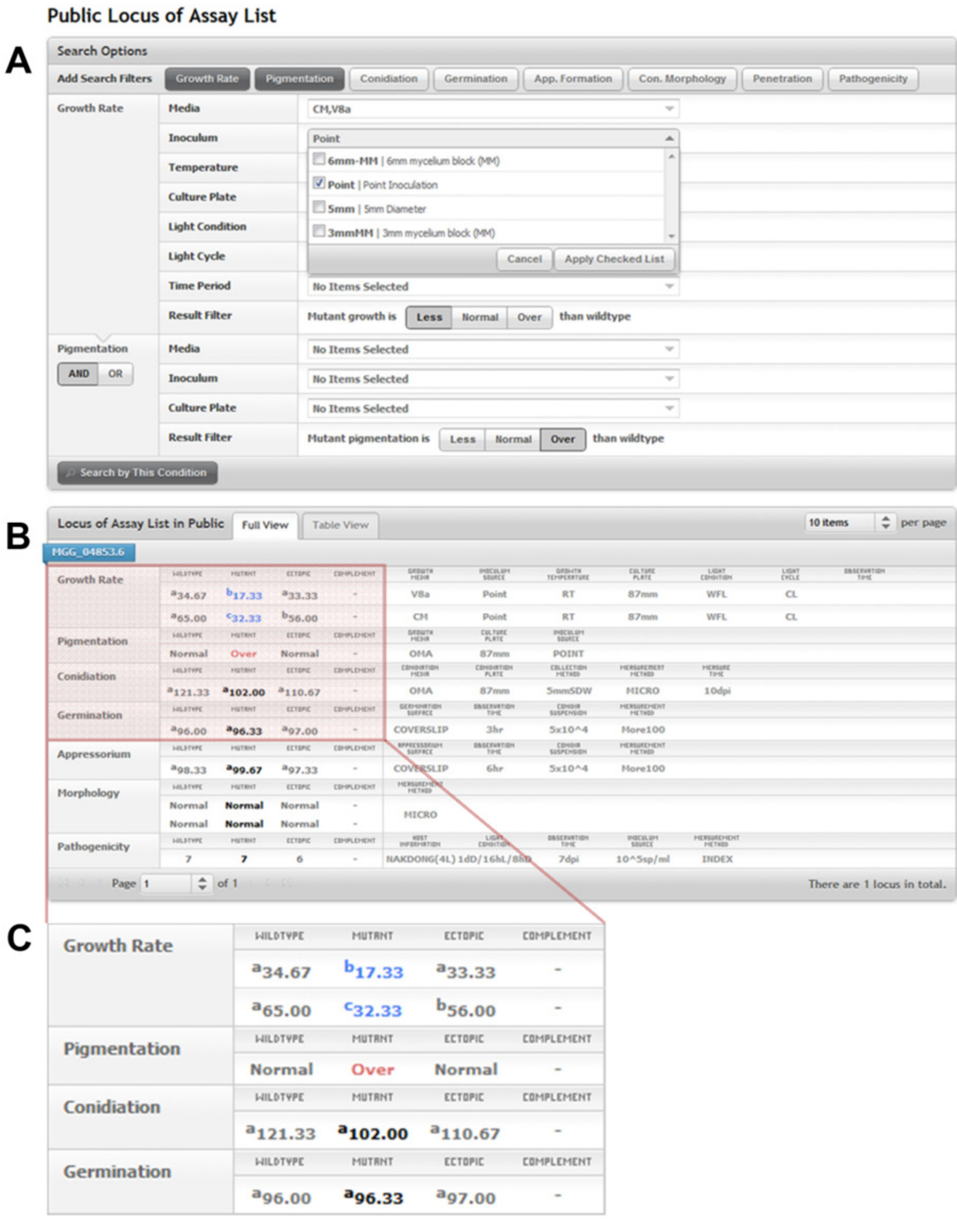
Supplement Figure S1. Data acquisition of phenotype assay in published paper. First, searching by keyword “Magnaporthe”, “Riceblast” and manually collect “Gene

Characterization" related. In that result we can get 124 papers. And filtered 73 papers they have sequence information about characterized gene. If they give only gene sequence, we found locus number mapping to *Magnaporthe oryzae* ver.6 genome by BLAST search. Finally, locus tagged all 67 genes in 57 papers that they assayed phenotype comparing wildtype between mutant are categorized by phenotype assay.



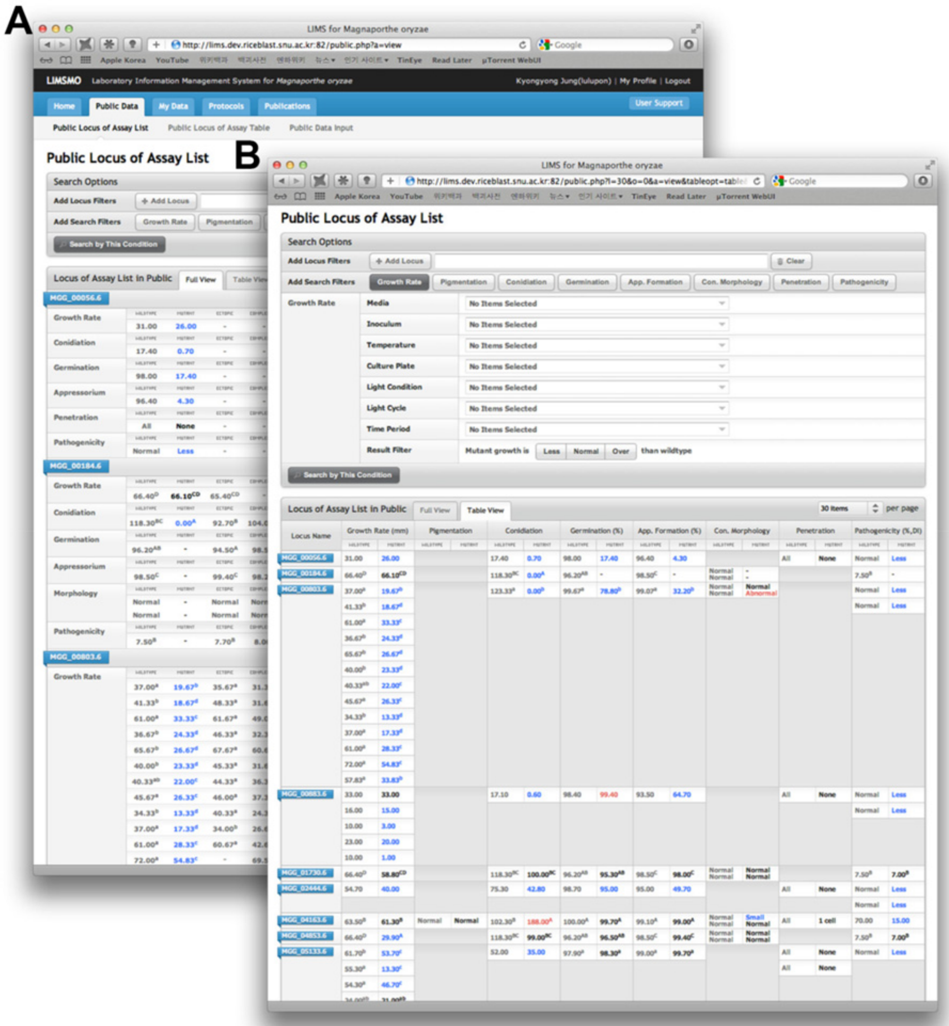
Supplement Figure S2. Scheduled procedure to phenotype experiment assay for efficiently managed a process of assay. Abbreviation means GR:Growth Rate, PG:Pigmentation, CN:Conidiation, GM:Germination, AP:Appressorium Formation, CM:Conidial Morphology, PN:Penetration, PT:Pathogenicity.

Supplement Figure S3. Front page of LIMSMO homepage. There are new project link button (Get Started) and browse published data link button (Browse Data). Side of content description, users can check the overall statistics like number of projects, locus of assay, protocols and stored papers.

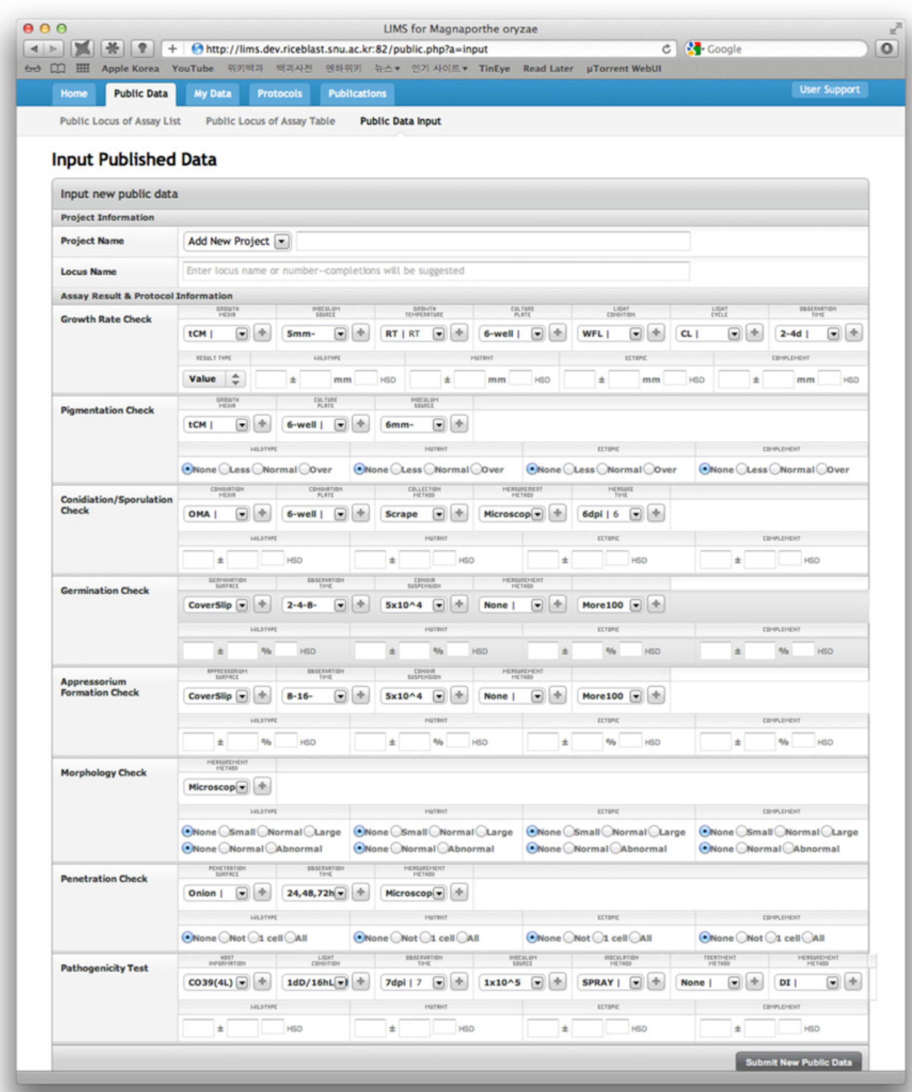


Supplement Figure S4. List of all locus from phenotype assay data that its public flag is published. (a) Users are able to combine any type of assay, protocols and assay

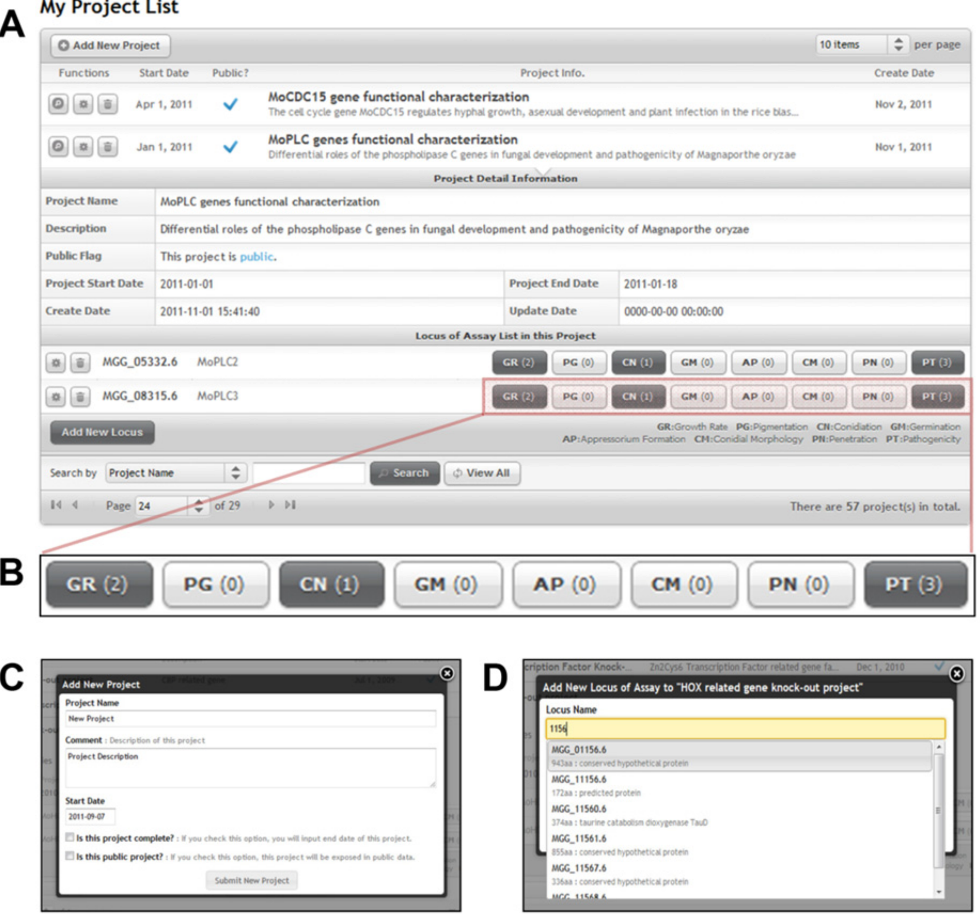
result by and/or logic. (b) Result and protocols of locus which is filtered by above filter. (c) Phenotype assay results are represented by color. Red color means over/large status, and Blue color means less/small status. And superscript alphabet on result is represent the group of Tukey's HSD Post-Hoc calculation. (Siginificance level is 95%)



Supplement Figure S5. Two way of view type for published data list. (a) In full view, users can see not only all phenotype assay result for each locus but also protocol information for assay. (b) In table view, it provides simple information about phenotype assay result for wildtype and mutant.

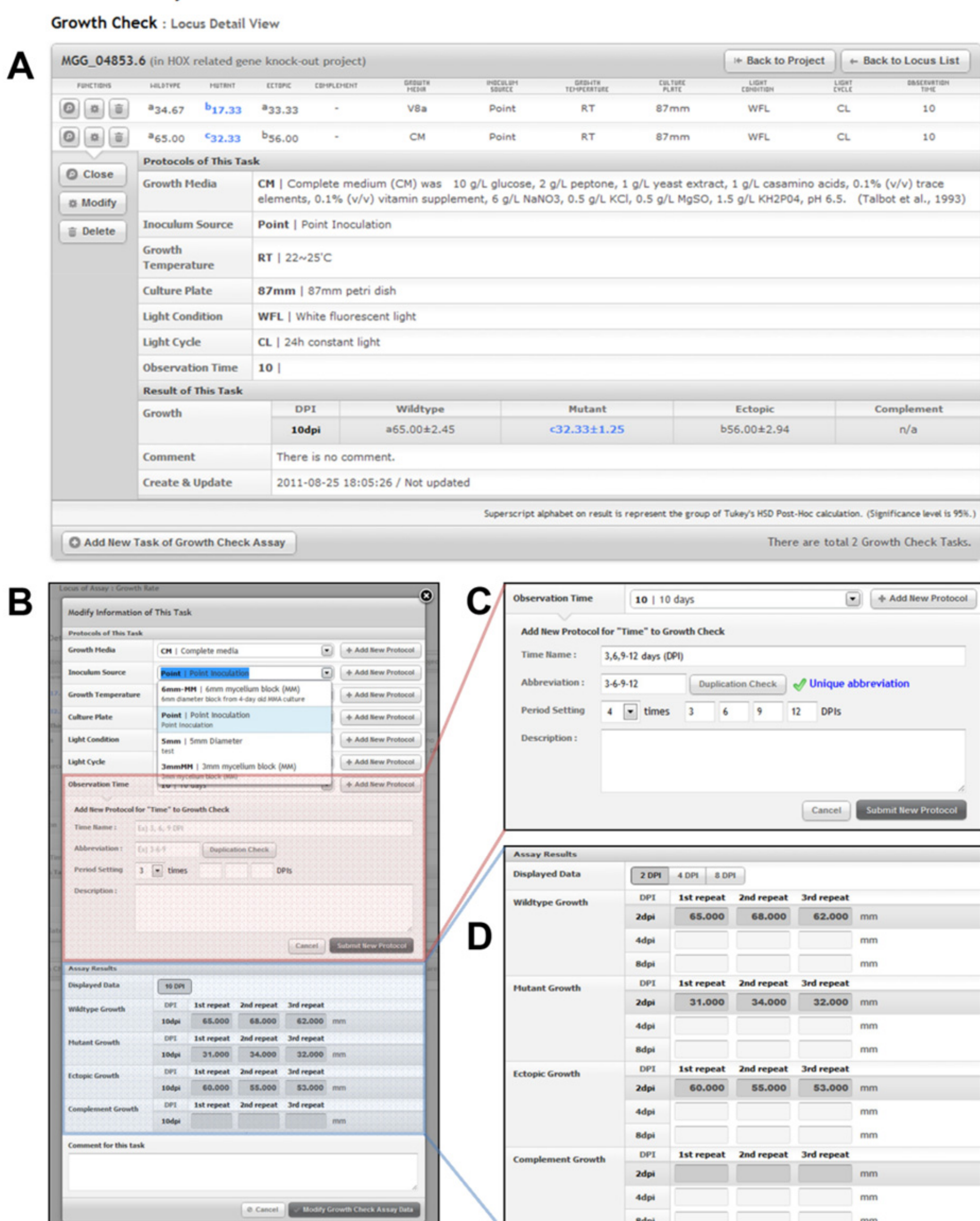


Supplement Figure S6. Published data input form. In these form, there are all eight phenotype assay result form. Researchers can input new or published data at once. And this method has no option for calculation of stastics (Mean, Tukey HSD).



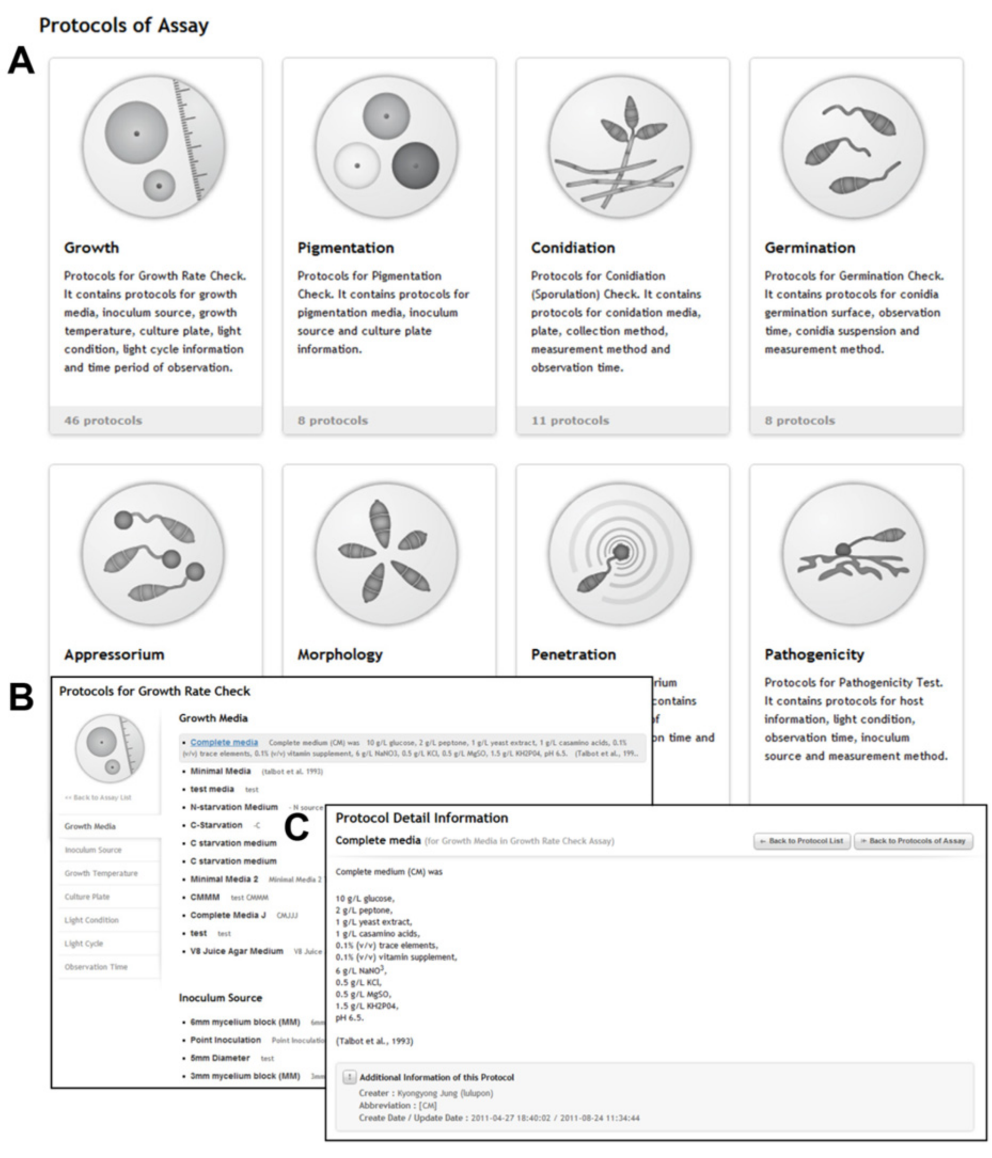
Supplement Figure S7. Project management page in My Data section. (a) List of all project that created by private users. In detail section, it shows locus information, progress status of 8-phenotypes assay. (b) In assay progress stuatus, abbreviation means - **GR**:Growth Rate, **PG**:Pigmentation, **CN**:Conidiation, **GM**:Germination, **AP**:Appressorium Formation, **CM**:Conidial Morphology, **PN**:Penetration, **PT**:Pathogenicity. Numbers on right side in that button represent the number of task in this assay. (c) Researchers can make new project or modify some project. Project

name, date of project start is required form. Additionally, user is able to select the opiton of public project flag and complete project flag. (d) Add new locus to phenotype assay into selected project. Input form of locus name has the function of auto complete.



Supplement Figure S8. Locus of Assay task management page in "My Data" section. (a) List of all task of phenotype assay in selected locus of assay. It shows result data

of each task and protocols that used in each task. (b) Add or modify some task function is separated two large section. Select or add new protocols for task is first section. And second section is input result data of each task. (c) User can make new protocol instantly in task input process. That figure shows add new protocol of observation time. It consist protocol name, abbreviation, time period setting(If it is required) and protocol description. (d) Input data of result form is composed three times repeat form. Assay of growth rate check, germination and appressorium formation check have time period option for observation. Thus, input form has selection of displayed data in that assay.

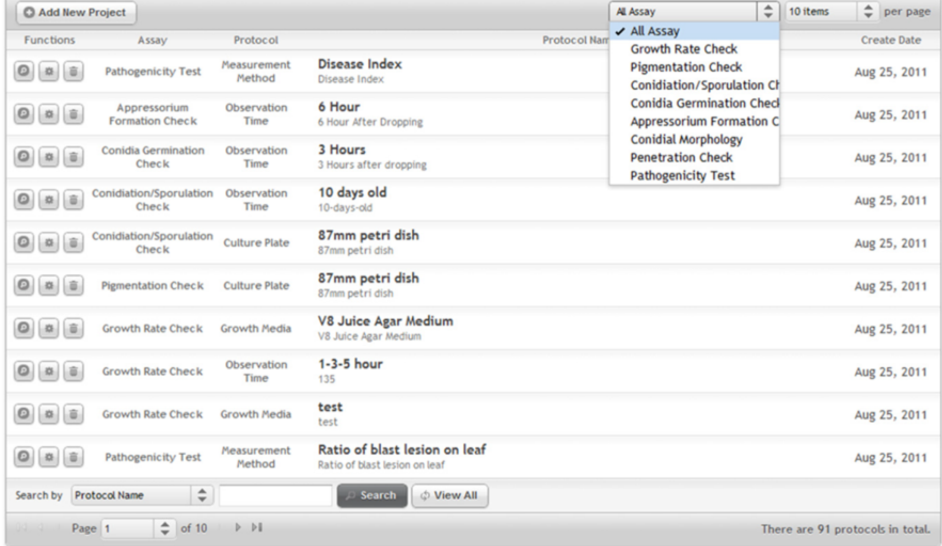


Supplement Figure S9. Protocols of assay page has the list of 8-group that categorized by phenotype assay. (a) There are eight groups of protocols for phenotype assay. (b) Protocols of assay inside, there is sub categorized menu that seperated by experimental instrument or method. And protocols are also splited by same manner.

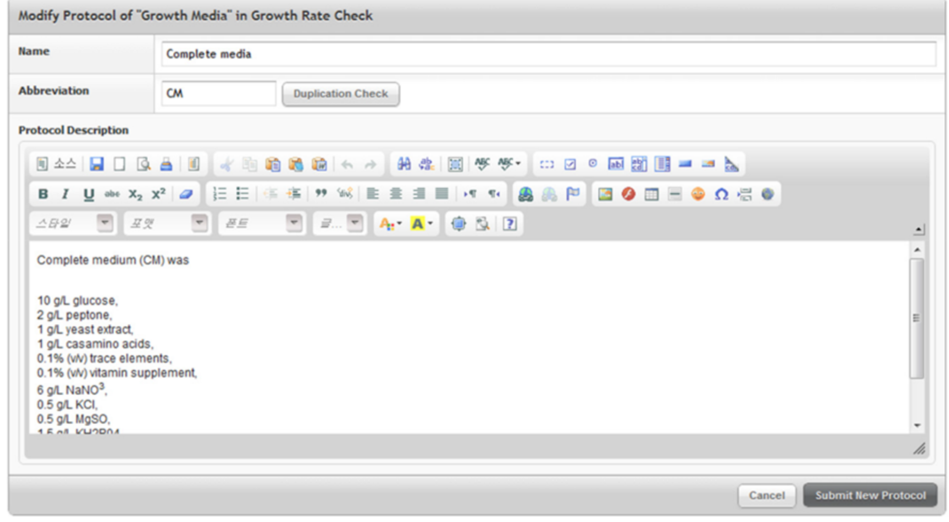
(c) Detail information of selected protocols has description of protocol and additional information. (creater, abbreviation, create and update time)

A My Protocols

Protocol List

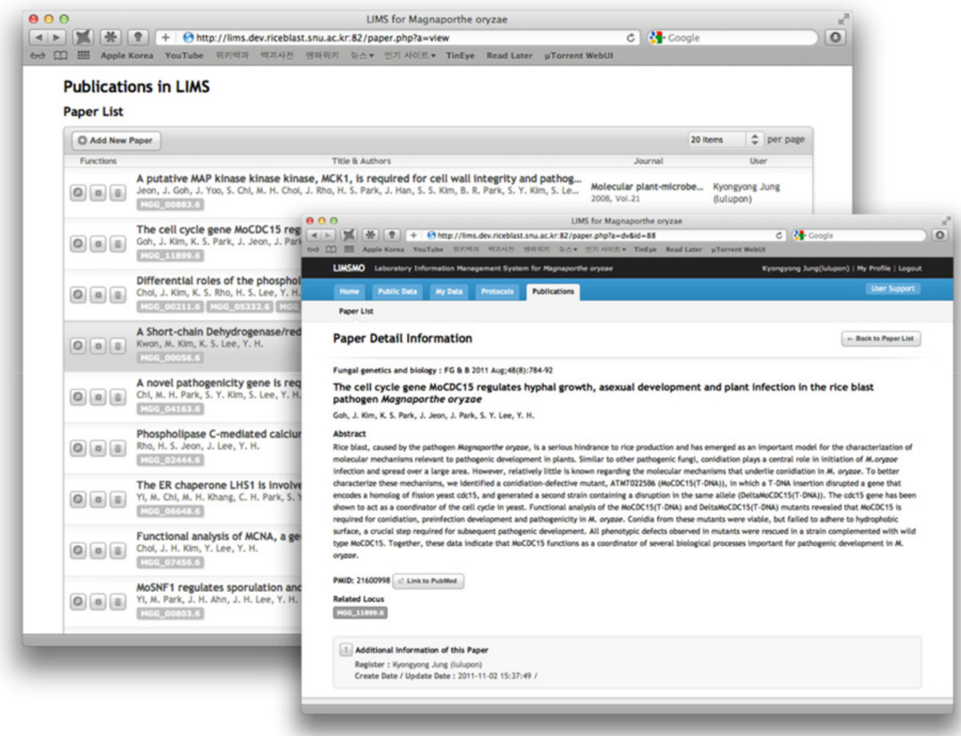


B Modify Protocol



Supplement Figure S10. My protocols management page in "Protocol" section. (a) List of all protocols that owned by user who create this protocol. It is able to filtered by phenotype assay and search by protocol name. (b) Add new or modify own

protocol. If the protocol is already used other task, user can not delete or modify entirely section of protocol. But in this situation, modified protocol is added after to old protocol. And It can not deleted but protocol is closed instead of deletion.



Supplement Figure S11. Publications in LIMSMO. This database also has all publications related functional characterization of genes. Any locus in CFGP data warehouse is able to linked for specific papers.

곰팡이 기능 및 비교 유전체학을 위한 유전체 탐색기와 실험 정보 관리 시스템

정 경 용

초 록

효모균 *Sacchromyces cerevisiae* 의 전체염기서열이 1996 년에 발표된 이래, 90 종 이상의 진균 염기서열이 공개되었다. 하지만 각각 다른 기관에서 진행하여 보관중인 염기 서열의 불균등성으로 인해 효율적이고 종합적인 비교분석을 위한 데이터의 통합이 방해되고 있었다. Comparative Fungal Genomics Platform 은 단일 표준화된 형식을 통하여 유전자 데이터의 다각적인 통합 분석을 지원하기 위해 개발되었다. 이러한 CFGP 및 관련 데이터베이스의 시스템 구성을 기반으로 종간에 걸친 효율적인 데이터 시각화 및 활용을 촉진하기 위해 새로운 유전체 탐색기가 필요했다.

그에 따라 개발된 SNUGB 는 91 종 (129 집합) 의 곰팡이, 33 종 (37 집합) 의 식물과 동물의 유전체 염기서열 정보를 담고 있으며 각 유전체의 속성 정보를 시각화, 비교 분석 가능한 플랫폼이다. 저장된 유전체 정보는 도해

다이어그램과 표, 일반 텍스트 방식의 3 가지 시각화와 6 가지 표시 선택사항을 통해 탐색할 수 있으며 분류학에 기초를 둔 계층구조 방식의 새로운 탐색기를 통하여 각 종의 정보를 빠르게 접근 가능하다. 그리고 SNUGB 내부에 포함된 BLAST annotation 프로그램을 통하여 다른 종과의 비교분석이 가능한 기능도 제공하고 있다. 또한 SNUGB 의 모듈형 설계는 다른 유전체 비교 분석 플랫폼에 쉽게 이식이 가능하고 지속적인 확장이 가능한 장점을 내포하고 있다.

이러한 SNUGB 는 곰팡이계(界) 뿐만이 아니라 다른 계와의 유전체 비교 기능 분석에 강력한 지원 시스템이 될 것이다.

또한 다량의 유전체 염기서열정보와 기능 유전체학이 발전함에 따라 대규모 표현형 검사 실험은 효율적인 데이터 처리라는 문제에 직면하고 있다. 이러한 연유로 여러 모델 생물의 표현형 검사 데이터 관리 및 검색을 지원하기 위해 실험실 정보 관리 시스템 (LIMS) 이 사용되고 있다.

하지만 미생물 표현형 연구에 LIMS 를 적용하려는 많은 노력에도 불구하고, 각각의 실험체의 표현형 정보를 얻기위한 지침이 서로 다르기 때문에 얻어진 데이터 또한 비교 분석이 힘들었다. 전 세계적으로 쌀 재배 지역에 걸쳐 가장 파괴적인 질병인 벼 도열병을 발생시키는 곰팡이인 *Magnaporthe oryzae* 를 위해 개발된 LIMS 플랫폼 또한 존재 하지 않았다.

이러한 문제를 해결하기 위해 개발된 LIMS for *M. oryzae* 는 표적 유전자 제거에 대한 통합 관리 시스템뿐만 아니라 실험 과정 관리 시스템과 같은

표현형 분석 및 데이터 수집에 관한 표준화된 지침을 연구자에게 주고 있다. 또한 앞서 개발된 CFGP 플랫폼과의 연동을 통하여 유전체 정보와 표현형 정보의 연결 고리를 제공한다.

이 시스템은 유전자 정보와 그 유전자에 관한 표현형 실험 절차, 표준화된 실험 방법, 관련된 출판 논문을 통합 관리하며 저장된 결과는 다른 연구자와 쉽게 공유 가능하다. 또한 새로운 유전자의 기능 실험 전에 기 저장된 유전자 표현형 실험 결과를 사용자가 원하는 방식의 검색을 통해서 미리 기능 예측이 가능하다. 이러한 모든 기능은 사용자 친화적인 웹 기반의 사용자 중심 인터페이스를 통해 쉽게 접근할 수 있다.

이러한 LIMS for *M. oryzae* 시스템은 벼 도열병에 대한 중요한 연구 자원이 될 것이며, 도열병뿐만이 아니라 다른 모든 식물병원성 곰팡이의 시스템 수준에서 질병의 기작을 이해하는데 큰 도움을 줄 것이라 예측된다.

주요어: 기능 유전체학, 벼 도열병균, 유전체탐색기, 병원성유전체학, 실험정보관리시스템, 시스템 생물학

학번: 2006-21477

감사의 글

소년이로학난성(少年易老學難成) 일촌광음불가경(一寸光陰不可輕), “소년은 늙기쉽고 학문은 이루기 어려우니 찰나의 시간일지라도 가벼이 여기지 말아라” 라는 뜻을 가지고 있는 주자의 권학문에 나오는 시의 첫 구절입니다. 스물여섯살에 대학원에 입학해 식물균병학연구실 구성원이 되어 보낸 7 년 반의 시간을 제가 과연 얼마나 유익하게 썼는지, 뜻했던 학문의 성취는 얼마나 이루었는지 졸업을 앞두고서야 이제사 한번 돌아보는 시간을 가지게 됩니다.

치기어린 마음에 응용생물학과 컴퓨터공학 복수전공을 시작했지만 제가 생각하던 이상과 현실이 달라서 고민하고 방황하던 와중에 저에게 두 학문의 접점이라 할 수 있는 생물정보학의 세계를 알려주고 이끌어주신 지도교수 이용환 선생님께 가장 먼저 감사의 인사를 드립니다. 학생보다 더 열정적이고 적극적인 모습을 보여주시며 자칫 지칠수도 있는 길고 험준한 학문의 길에서 언제나 밝은 빛과 올바른 길을 제시해주시는 선생님의 모습이 제 대학원 생활에 있어 가장 큰 힘이었다 생각합니다.

박사 학위가 끝이 아니라 그 후의 더 넓고도 심오한 학문의 세계에 대해서 조언을 해주신 심사위원장 최도일 교수님, 김희발 교수님, 김지현 교수님, 박숙영 교수님께도 아직 부족한 점이 많은 저를 인정하고 격려해주셔서 감사하다는 말씀을 올리고 싶습니다. 또한 학부시절부터 대학원시절까지 전공에 대해 많은 가르침을 해주신 식물미생물학과의 이인원 교수님, 황인규 교수님, 가종억 교수님, 박은우 교수님, 김영호 교수님께도 학생으로써 머리숙여 감사드립니다.

연구실에 계신동안 학문적으로나 인간적으로 큰 버팀목이 되어 주셨던 김경수 교수님과 연구 방향에 대해 아낌없는 조언을 해주시고 심사위원까지 맡아주셨던 박숙영 교수님, 먼 미국에 계시지만 기꺼이 제 부족한 논문을 다듬어 주시고 격려해주신 강석찬 교수님께도 진심으로 감사의 말씀을 올립니다.

대학원 생활동안 함께 해준 균병학 연구실 구성원들에게도 이 자리를 빌어 감사드립니다. 처음 대학원생이 되어서 많은 가르침을 받았던 노희술 박사님, 실험실 생활에 있어 가까운 곳에서 큰 힘이 되어준 최진희 박사님과 박주영 박사님, 유쾌함과 열정으로 저를 비롯한 후배 대학원생들을 이끌어 주었던 지명환 박사님과 이미화 박사님, 늘 밝은 기운을 북돋아준 최재혁 박사님, 올바른 연구자의 자세를 제시해 준 전준현 박사님들에게 후배로써 존경의 의미를 담아 감사의 인사를 드립니다. 그리고 저와 지금까지 많은 시간을 동고동락한 든든하고 힘이되는 후배 성형이와 재진이, 활력소가 되고 있는 실험실의 젊은 피 현정, 수빈, 종범, 성범, 아람, 다영, 서문이, 먼 나라에 와서 수고하고 있는 사닷과 알벨리 한테도 고마움을 전합니다. 또한 연구에만 전념할 수 있도록 각종 힘든 일을 도맡아 해주시면서도 언제나 웃는 모습을 보여주신 윤여경 선생님과 졸업하고도 잊지않고 찾아와 주는 성용형, 민중형, 선영, 세은, 세련, 효정, 동한, 가은이 에게도 실험실 생활동안 받았던 도움에 감사드립니다.

무엇보다 FBL 멤버들에게도 감사의 말을 전하고 싶습니다. 이제는 든든한 버팀목이 되어준 후배지만 많은 것을 배우게 되는 재영이, 다방면으로 도움을 받은 경채, 동갑내기로 늘 힘이 되어준 길원이, 연구실의 파릇파릇한 새싹 기태,

실험실은 다르지만 많은 조언을 해주신 김용민 박사님과 승일이가 있어서 제 대학원 생활에 큰 힘이 되었습니다.

사회생활에 있을 부족함을 메꿔준 인생 선배이자 기쁠때나 슬플때나 함께해준 친구처럼 지내온 호성님, 항경님, 성기님, 동현님께도 이 자리를 빌어 감사의 말씀을 드립니다.

긴 대학원 생활을 하는데 있어서 가족들의 격려와 성원이 없었다면 제가 이자리에 있을 수 없었을 것입니다. 부족한 아들을 언제나 사랑으로 감싸고 이끌어주신 부모님, 먼 미국땅에 있지만 항상 마음으로 통하는 나의 하나밖에 없는 형제 보람이, 멀리 떨어져 계신 부모님의 빈 자리를 메꿔주신 사랑하는 나의 할머니, 조카를 물심양면으로 챙겨준 고모들께 손자이자, 아들이자, 조카이자, 형으로써 고마움을 전합니다. 그리고 이십대 초반부터 어떤상황에서도 언제나 제 곁에 있어주었고 지금은 영원한 일생의 반려자가 된 무엇보다 소중한 사랑하는 나의 아내 효진이에게도 진심으로 감사의 마음을 전합니다.

제 주변 많은 분들의 관심과 격려로 인해 지금 이 자리까지 왔습니다. 지금까지 겪었던 모든 경험을 바탕삼아 앞으로 더 나은 모습을 보여줄 수 있도록 더욱 분발하고 노력하겠습니다. 다시 한번 감사드립니다.

2013 년 8 월

정경용 올림