Bioinformatics platform for ICRISAT's global research needs

Jayashree B, Chandra S, Hoisington DA, Upadhyaya HD, Hash CT, Vadez V, Varshney RK, Senthilvel S.

Background

Scientists in ICRISAT’s Global Theme on Biotechnology employ a range of modern genomic technologies in their efforts to enhance the efficiency and effectiveness of crop improvement. These include high-throughput genotyping technologies. The rate-limiting step in genomics is no longer data generation but rather the speed, at which data is captured, validated, analyzed and turned into useful knowledge. The role of Bioinformatics is to remove this rate-limiting step through the development of software platforms for handling large volumes of data generated and facilitate its analysis. The term ‘platform’ implies all those vital services and technologies that are needed to support genomics research projects at all of ICRISAT locations. Bioinformatics at ICRISAT functions both as a research and research support unit.

Research efforts in Bioinformatics at ICRISAT focus on two major areas. One relates to infrastructure development which includes: (a) the development of appropriate software and efficient protocols for data capture, storage, retrieval and dissemination; validating the large volumes of data with special emphasis on data quality and (b) the development of analysis tools such as integrated decision support systems for scientists and software pipelines to accelerate research efforts in molecular marker discovery, annotation and comparative genomics. The other area relates to data analysis and algorithm development relevant to comparative biology and specific to cereal – legume relationships.

The Bioinformatics support offered to the scientific community are need based. Support provided is in areas such as generation of databases, programming, web interfaces and LIMS for genotyping workflow management. We also provide support with bioinformatics analysis tools, phylogenetic analysis, sequence annotation, batch web submissions and downloads and installation support for bioinformatics freeware.

The hardware infrastructure in Bioinformatics has been steadily growing since 2004. We now have one middle level server, one Paracel Linux cluster consisting of 4 dual AMD Opteron processors (for high-end computing jobs) and 16 Pentium IV PCs. The open source movement and its philosophy have largely driven platform software development in Bioinformatics at ICRISAT. This is because programming with open source has several advantages: the absence of licenses, lack of hardware and operating system dependencies, better performance in large scale programmability or collaborative development environments. The more important advantages are in terms of ownership and extensibility. Almost all the software development is being done using the freely available Perl or Java programming languages. Since academic and research institutions have taken to the open source culture there is a number of bioinformatics software available in the public domain. All the data analysis pipelines constructed at ICRISAT has been built upon freely available software. All tools developed here are put back into the public domain for others to use and extend.

Activities

The activities have been grouped under four sections: infrastructure development, research in the area of comparative genomics, capacity building and linkages and partnerships.

Infrastructure development:

(a) We have developed a Laboratory Information Management System (LIMS) between the years 2004-2006. This LIMS meets the needs of a moderately high-throughput molecular genotyping facility. The beta testing of the LIMS system is in progress at ICRISAT. The data producers have uploaded genotyping data into the system for the composite core collections of chickpea, sorghum and pearl millet. The application functionality and user interfaces continue to be modified to suit user requirements. With the addition of a few functional modules, the application was successfully transferred to BeCA (Biosciences Eastern and Central Africa facility) at Nairobi and IITA-Ibadan. We have also developed an Inventory management system, a web based application widely used in the genomics laboratory.

(b) The ICRIS database (Integrated ICRISAT Crop Resources Information System) is being developed to integrate genotyping information with genetic resources and phenotype information. The LIMS –ICRIS software adaptor allows the flow of genotyping data from the LIMS application into the ICRIS database. The database implements the open standards advocated in the GCP (Generation Challenge Program) software platform. This will allow for the database to become interoperable with other databases and repositories available in the public domain. Interfaces are also being written that will allow the access of data within the database through publicly available analytical or visualization tools.

(c) We have developed an integrated decision support system, called iMAS, to seamlessly facilitate marker-assisted plant breeding by integrating freely available quality software involved in the journey from phenotyping-and-genotyping of genetic entities to the identification and application of trait-linked markers, and providing simple-to-understand-and-use online decision guidelines to correctly use these software, interpret and use their outputs. Into its third year of development this software is being tested by users from national institutes as well as international collaborators and continues to be improved for use as a standalone application.

(d) High performance computing toolbox: The Paracel Linux cluster hosts pipelines (a series of different software through which data is pipelined) and standalone software analysis tools relevant to comparative genomics and population genetics. The comparative genomics toolbox now includes a suite of programs that have been implemented to work in parallel (on all 4 nodes of the cluster). This was possible through tapping the parallel programming expertise available with the Advanced Technology Centre at TCS (see Linkages and Partnerships below). These programs are useful for marker mining from large public datasets, comparative sequence analysis and phylogeny. The population genetics toolkit includes a parallelized version of the program ‘structure’ with user interfaces and visualization software along with format conversion tools.

(e) Two molecular marker databases have been published online, both constructed from public sequence data and relevant to ICRISAT mandate crops: an SSR (simple sequence repeats) database (http://www.intranet.icrisat.org/gt1/SSR/SSRdatabase.html) and a database of tentative orthologous groups specifically derived from stress transcripts (http://www.intranet.icrisat.org/gt1/tog/homepage.html).

In-silico comparative genomics: Research activities pursued in this area have a direct bearing on genomics projects and have led to several joint publications with laboratory staff. Activities include the evaluation of algorithms for orthology detection and phylogenomic studies of protein family orthologs related to the plant abiotic stress response.

Capacity building: Over the past few years, there has been increased interest to work in ICRISAT's Bioinformatics Unit by students from universities and institutions in India, almost all projects are in the area of comparative biology and phylogenomics. During the last two years, some ten graduate students have been accommodated in training programmes each year. These students are working towards a degree in biotechnology or bioinformatics and the programme provides them with hands-on training opportunities. In-house capacity is also being strengthened through working with collaborators from Central Institutes like the Centre for DNA Fingerprinting and Diagnostics and International Institute of Information Technology, both located in Hyderabad.

Linkages and partnerships: ICRISAT cannot do it alone. So we often seek out consultants and third parties that can provide the required expertise in bioinformatics. One such company is the Tata Consultancy Services (TCS), a leading global information technology consulting, services and business process outsourcing organization. Given our successful sub-contracting of development of parallelized software for the high performance computer at ICRISAT, we have wanted to extend our linkages with companies like these. We have been working with scientists of the Advanced Technology Centre, the R&D wing for bioinformatics at TCS, to build partnerships in the areas of software platform development, high-throughput comparative biology and systems biology.

Highlights of achievements:

The **LIMS** application received a boost when the Bioinformatics Unit at CIMMYT – a sister CGIAR center – conducted an evaluation of various LIMS applications in early 2006 and regarded ICRISAT's LIMS to be superior. Since then the application has been downloaded by 186 different users in India and abroad with a total of 486 downloads recorded from the project page (http://www.icrisat.org/gt-bt/lims/lims.asp). The French public institute, CIRAD, and the Brazilian national program, EMBRAPA have evinced interest in using and contributing to the growth of this application. The LIMS in the year 2007 is moving towards collaborative development with the fostering of a community of software developers from different CG institutions. This team of developers will further extend the ICRISAT LIMS for implementation within their institutions.

Data validation and quality: A major concern when working with large volumes of data is the risk that errors are introduced at various levels of data generation, entry and manipulation. The LIMS helps to reduce these at the data capture point by validating data entry, checking for inconsistencies and annotating data at every step of the workflow. This assures that the raw data captured is maximally accurate. A critical next step is often some form of data manipulation. For SSR genotyping using modern DNA sequencers, the manipulation step is to turn a "raw" allele size into its "actual" discrete size. Algorithms that provide a strong statistical basis for these "allele calls" have been implemented in a software package "**Allelobin**". The software is available both within the LIMS package, and as a standalone application. Allelobin has been widely distributed at workshops and is being used by scientists at CIRAD, IPGRI and Kasetsart University in Thailand.

The **iMAS** standalone software is finding considerable interest amongst the NARS community. Two major workshops were conducted during 2006 accommodating over 40 scientists from the NARS besides ICRISAT staff and interested private partners to test the application. The application will be released formally in September 2007.

Looking forward: The outputs and outcomes scheduled and reported in the medium term plan are the milestones that we look forward to achieving in the next two-three year period (2007-2009). This includes advancing development and testing of information systems for the management, analysis and interpretation of genomics data and making these globally accessible and available. This will include the integrated database for genotyping and phenotyping data, besides tools for marker development, marker assisted selection and marker accelerated backcross for ICRISAT mandate crops. We expect that the setting up of the Centre of Excellence in Genomics (CEG) will increase the importance of the role that Bioinformatics will play in providing support to the very high throughput data generating technologies being implemented at the Genomics laboratory at ICRISAT.

For additional information/clarification, contact Dr Jayashree, B (b.jayashree@cgiar.org)

Figure 1: A few user interfaces from the LIMS

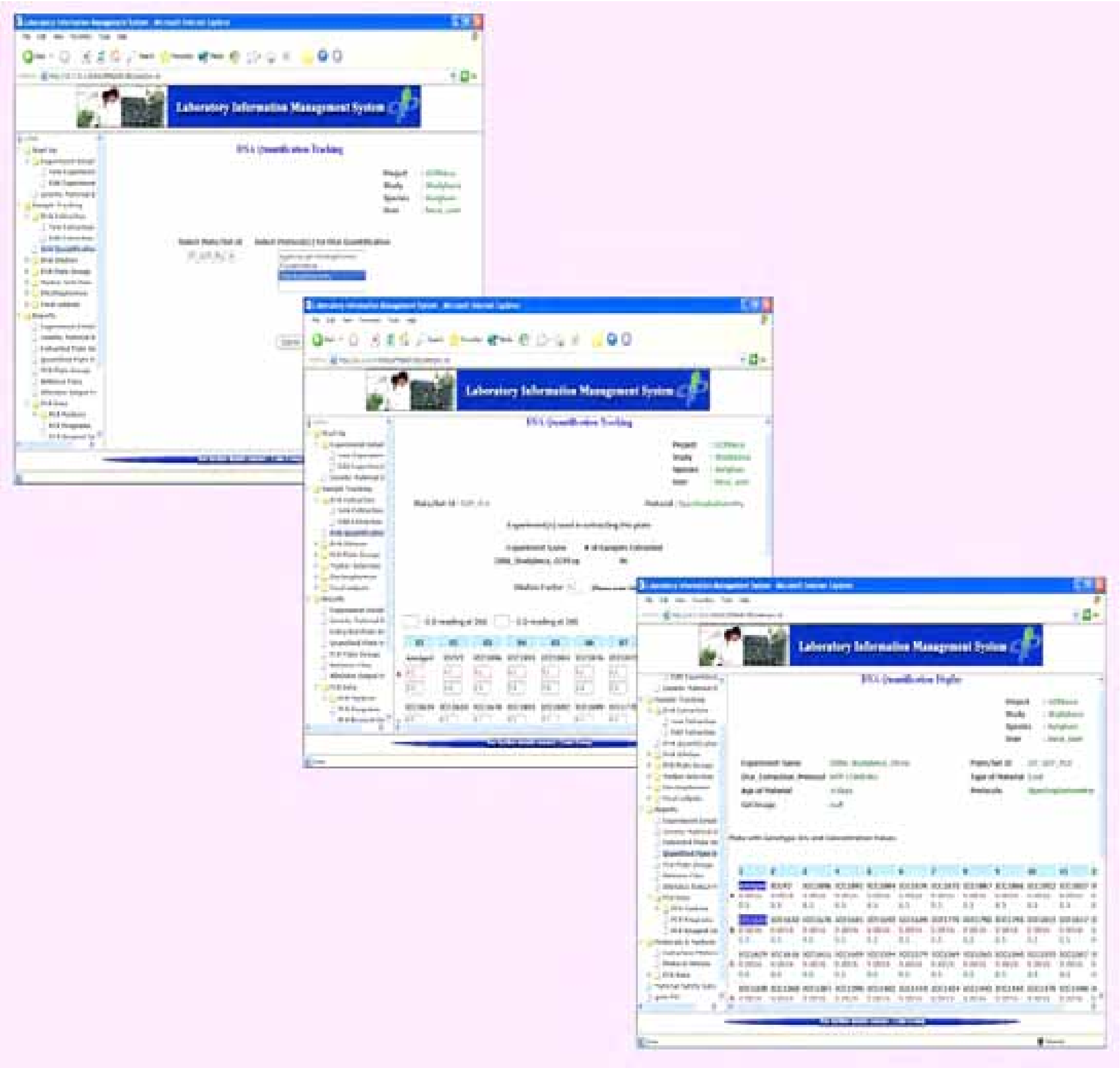
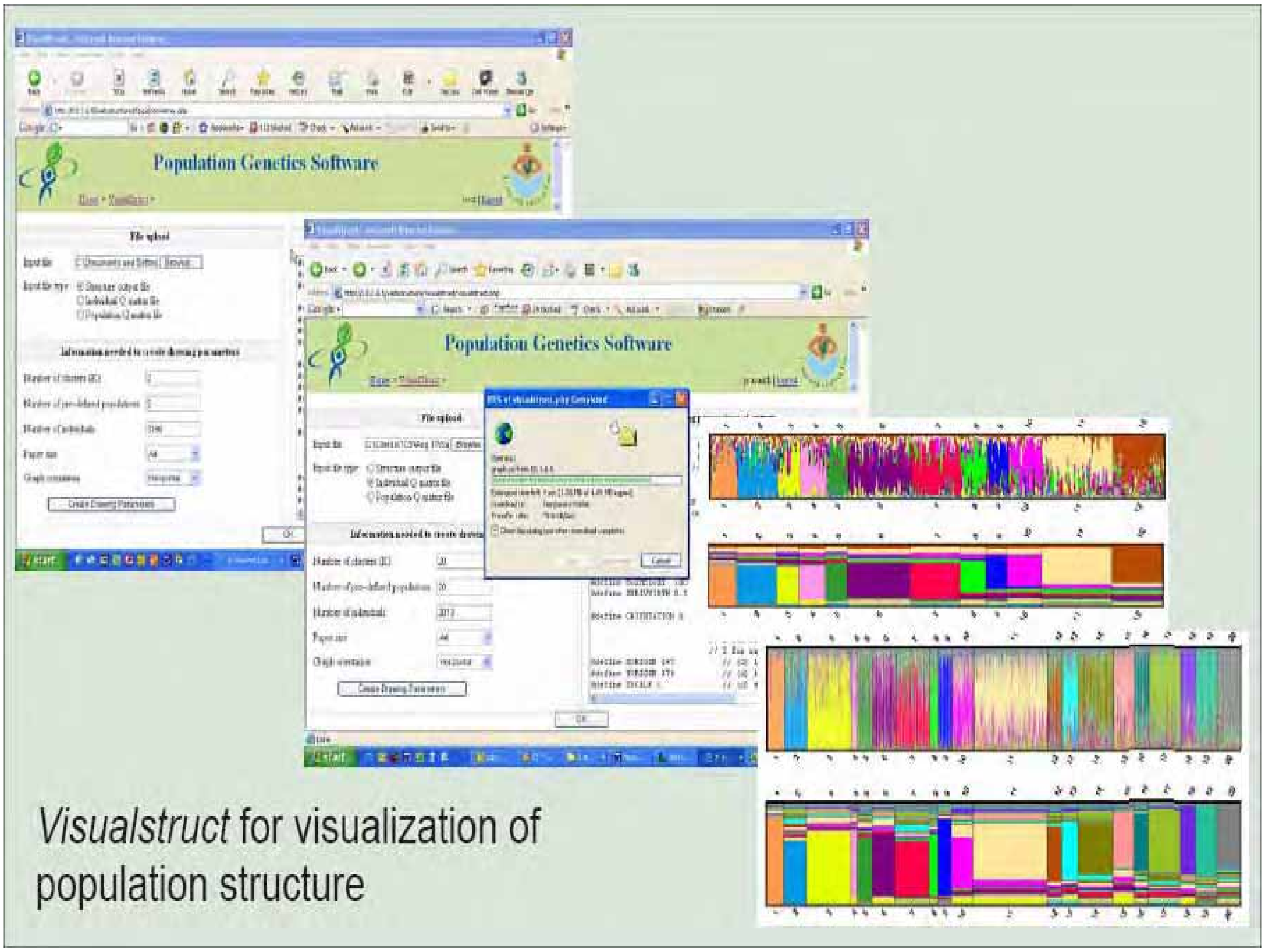

Laboratory Information Management System
Laboratory Information Management System
Laboratory Information Management System

Figure 2: The interfaces to the software "Structure" and visualization tool on the Paracel Linux cluster.



Population Genetics Software

File upload

Input file

Input file type: ○ Structure output file
○ Individual Q matrix file
○ Population Q matrix file

Information needed to create drawing parameters

Number of clusters (K): 20

Number of pre-defined populations: 20

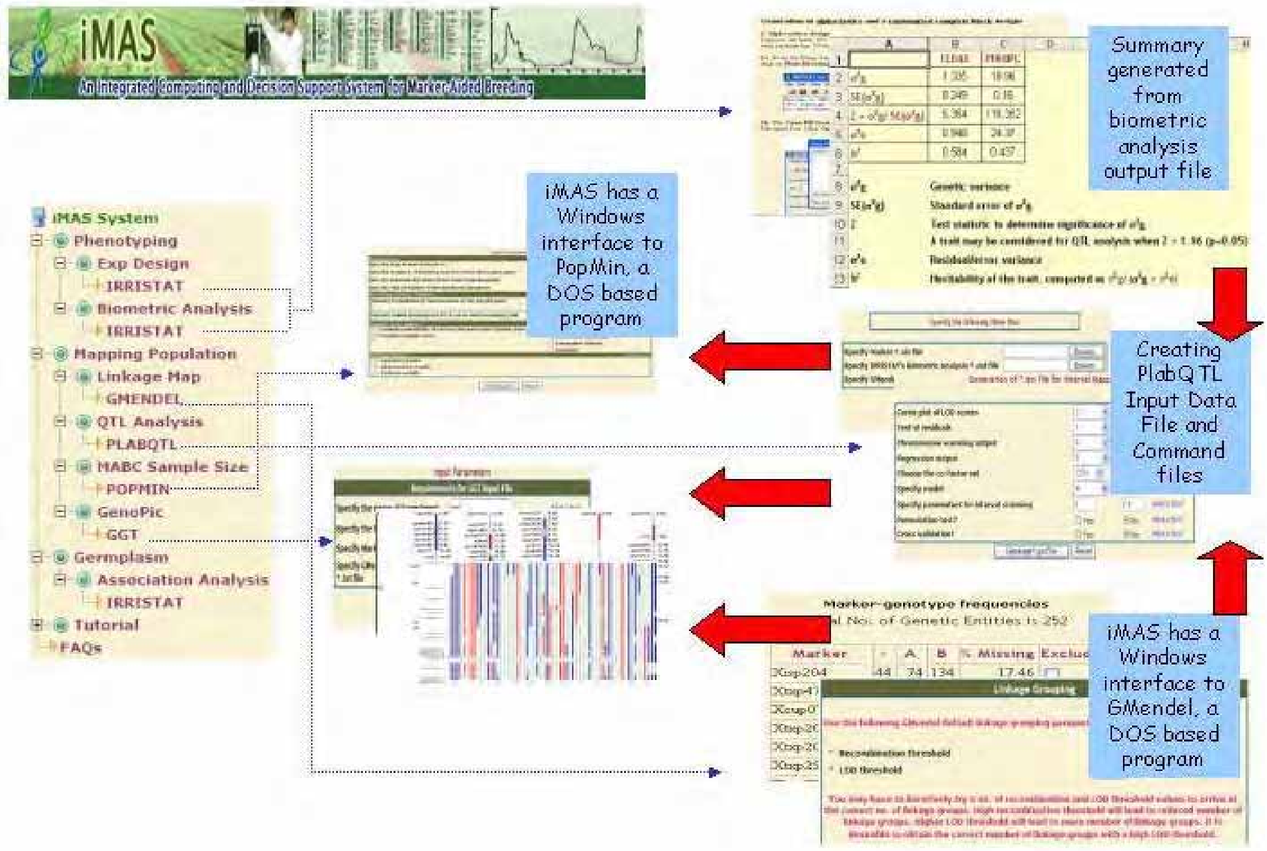
Number of individuals

Paper size: A4

Graph orientation: Horizontal

Visualstruct for visualization of population structure

Fig.3: The interfaces in the standalone implementation of the “iMAS” software.



Summary generated from biometric analysis output file

iMAS has a Windows interface to PopMin, a DOS based program

Creating PlabQTL Input Data File and Command files

iMAS has a Windows interface to GMendel, a DOS based program