4th STATSEQ Workshop

18th and 19th April, 2012

Polo Zanotto
University of Verona (Italy)

INDEX

SCIENTIFIC COMMITTEE

Thomas Schiex
Brigitte Mangin
Alberto Ferrarini
Marco Bink

LOCAL HOST

Università degli Studi di Verona
Prof. Massimo Delledonne

LOCAL ORGANIZING COMMITTEE

Anne-Marie Digby
Alberto Ferrarini
Luca Venturini
Andrea Minio

ORGANIZING SECRETARIAT

COGEST M. & C. SRL
Vicolo San Silvestro, 6 - 37122 Verona (I)
Tel. +39 045 597940 - Fax +39 045 597265
e-mail: cogest@tin.it - www.cogest.info

VENUE

The Workshop will be held at the Lecture Room Aula Magna T2 of Polo Zanotto located at Viale Università 2 of the University of Verona, Italy (see map at page 57)

Wednesday 18th April 2012

08h30		StatSeq Workshop Registration opens	
09h20	*10′*	***Marco Bink*** - Welcome & objectives COST Action & workshop	
		SESSION 1: CHAIR MASSIMO DELLEDONNE	
09h30	*40′*	Invited speaker	**Michele Morgante** *University of Udine* From one for all to all for one: the NGS revolution in plant science research
10h10	*20′*	Contributing speaker	**Alberto Ferrarini** *University of Verona* Assembly and temporal characterization of the V. vinifera cv. Corvina berry transcriptome
10h30	*30′*	*Coffee break*	
		SESSION 2: CHAIR JORG BECKER	
11h	*25′*	Invited speaker	**Dan Bolser** *Wellcome Trust Genome Campus, UK* TransPLANT, developing a "trans-National Infrastructure for Plant Genomic Science"
11h25	*20′*	Contributing speaker	**Berkhard Linke** *Center for Biotechnology, Bielefeld University, Germany* NGS data processing with Conveyor workflows
11h45	*20′*	Contributing speaker	**Finn Drabløs** *Norwegian University of Science and Technology, Norway* MotifLab: A tools and data integration workbench for motif discovery and regulatory sequence analysis
12h05	*15′*	Posters	flash presentations (1 slide & 1 minute)
12h20	*90′*	*Lunch break & poster viewing*	
		SESSION 3: CHAIR HARALD MEIMBERG	
13h50	*40′*	Invited speaker	**Maria Colomé-Tatché** *Groningen Bioinformatics Centre, The Netherlands* Genome-wide methods for the study of DNA-methylation inheritance
14h30	*20′*	Contributing speaker	**Guillem Rigaill** *Bioinformatics and Statistics and Cancer Systems Biology Center, The Netherlands* A statistical approach to estimate copy number from NGS capture sequencing data
14h50	*30′*	*Coffee break*	
		SESSION 4: CHAIR DIMITAR VASSILEV	
15h20	*40′*	Invited speaker	**Lauren McIntyre** *Molecular Genetics and Microbiology, University of Florida* Variability in RNA-seq: design and modeling
16h	*20′*	Contributing speaker	**Julie Aubert** *AgroParisTech, France* A Comprehensive Evaluation of Normalization Methods for High-Throughput RNA Sequencing Data Analysis
16h20	*20′*	Contributing speaker	**Marie Laure Martin-Magniette** *INRA, France* Model-based clustering for high-throughput sequencing data to determine similar expression profiles across genes
16h40	*60′*	Round Table Discussion	Transcript Quantification from RNA seq data Co-chairs: **Pawel Krajewski & Marie-Laure Martin-Magniette**
17h40		*Drinks and poster viewing*	
19h00	*30′*	*Evening stroll to "Osteria da Ugo" dinner venue*	
19h30		Dinner at Osteria da Ugo	

Thursday 19th April 2012

08h30		StatSeq Workshop Re-opens	
		SESSION 5: CHAIR ANDREAS VOLOUDAKIS	
09h00	*40′*	Invited speaker	**Jonathan Marchini** *Department of Statistics, University of Oxford, UK* Haplotype estimation and imputation using low-coverage sequence data
09h40	*20′*	Contributing speaker	**Thomas Odong** *Wageningen University and Research, The Netherlands* SNP discovery from Next Generation Sequencing to study for genome variation in Arabis alpina natural populations
10h00	*20′*	Contributing speaker	**Jan de Boer** *Wageningen University laboratory of Plant Breeding, The Netherlands* Genotyping by sequencing tetraploid potato and attempts towards reconstruction of haplotypes
10h20	*30′*	*Coffee break*	
		SESSION 6: CHAIR SOREN BAK	
10h50	*40′*	Invited speaker	**Jeff Glaubitz** *Institute for Genomic Diversity, Cornell University, USA* Genotyping by sequencing in maize
11h30	*20′*	Contributing speaker	**Patrik Waldmann** *Thetastats, Sweden* The effect of LD between SNPs on some statistical methods for GWAS
11h50	*20′*	Contributing speaker	**Ron Wehrens** *Fondazione Edmund Mach, Italy* Meta-statistics for Biomarker Selection in the Omics Sciences
12h10	*20′*	Contributing speaker	**Willem Kruijer** *Wageningen University and Research Centre, The Netherlands* Sequential Monte Carlo algorithms for high throughput genetic mapping from dense genomewide SNPs
12h30	*90′*	*Lunch break & poster viewing*	
		SESSION 7: CHAIR REMY BRUGGMANN	
14h00	*40′*	Invited speaker	**Mark A DePristo** *The Broad Institute of MIT and Harvard, USA* Under the hood of the 1000 Genomes Project
14h40	*20′*	Contributing speaker	**Jimmy Vandel** *INRA, France* Towards Arabidopsis thaliana genetic regulatory network using discrete Bayesian network
15h00	*20′*	Contributing speaker	**Jaap Buntjer** *Keygene N.V, The Netherlands* Genomic Breeding using variomics data and decision support systems
15h20	*30′*	*Coffee break*	
15h50	*65′*	Round Table Discussion	Sequencing, genotyping & imputation strategies Co-chairs: **David Marshall & Marco Bink**
16h55	*5′*	M. Bink & T. Schiex	Concluding remarks
17h00			Workshop Closure

ORAL PRESENTATIONS

From one for all to all for one: the NGS revolution in plant science research

Morgante, M.[1,2,*]

[1]Dipartimento di Scienze Agrarie ed Ambientali, Università di Udine,
Via delle Scienze 208, 33100 Udine, Italy
[2]Istituto di Genomica Applicata, Parco Scientifico di Udine, Via J. Linussio 51, 33100 Udine, Italy

*Presenting author: **Michele Morgante** (michele.morgante@uniud.it)

The genomics revolution of the last 15 years has improved our understanding of the genetic make up of living organisms. Together with the achievements represented by complete genomic sequences for an increasing number of species, high throughput and parallel approaches are available for the analysis of DNA sequence variation, transcripts, proteins. The use of genomic tools has allowed us to start to unravel the genetic make up of traits that are relevant to plant breeding. At the same time a deeper understanding of what natural variation is at the sequence level has also been achieved, allowing us to realize that nature can sometime have much greater fantasy and inventiveness than any laboratory scientist and that genetic variation is continuously created in crop species. The pace at which we can analyze natural sequence variation has recently been greatly accelerated thanks to the advent of new DNA sequencing technologies and our ability to produce sequencing data is no longer a limiting factor and has greatly surpassed our scientific ability to provide a meaning to the sequences and their variations. This is best exemplified by the so called “missing heritability” problem in human genetics, where we can not provide a satisfactory explanation for the inheritance of complex traits, despite all the sequencing power that has been deployed. So now the ball is in the court of science that needs to make the next move and to develop models and methods to make full use of the huge amount of information that is produced daily in the laboratories around the world. The scientific revolution determined by NGS technologies is perhaps best exemplified by a change in the way we look at genomes, namely in going from the mean of many genomes (an overall view of what a genome is) to the many genomes (a detailed view of what each genome looks like) within a species and within an individual. We can finally examine individual events and not the mean of many different events. Entirely new possibilities are now open and it is now time for science to adventure into new areas and lead us to a better understanding of the functioning of living organisms. Proper and new statistical approaches are required for taking full advantage of these new opportunities.

Assembly and temporal characterization of the V. vinifera cv. Corvina berry transcriptome

Venturini, L.[1]; Ferrarini, A.[1]*; Minio, A.[1]; Buson, G.[1]; Pezzotti, M.[1]; Zenoni, S.[1]; Fasoli, M.[1]; and Delledonne, M.[1]

[1]Department of Biotechnology, University of Verona, Italy

*Presenting author: **Alberto Ferrarini** (alberto.ferrarini@univr.it)

The various strains of a species, especially in plants, harbor a huge sequence diversity at the genome level. In an age when the availability of sequencing trends steadily upward, this fact has important repercussions on the strategies we choose to analyze the different pan-genomes. Genomic resequencing for all strains of interest is a successful but time-consuming and expensive strategy. Moreover, the drawback of detection of variants based on genome-on-genome comparisons is that the balancing effect of mutually compensating mutations is not readily apparent. In contrast, the direct reconstruction and analysis of the expressed mRNAs in varieties can lead to a quick and accurate estimate of the effectual differences between transcriptomes.

We developed a pipeline to enact this strategy and employed it in the transcriptome analysis of a grape cultivar specific to the Verona area, Corvina. Initially we sequenced pooled cDNA isolated from 45 different conditions and organs. The different transcripts were assembled and later analyzed to discard contaminants and identify 228 sequences specific of our variety. They are enriched, in particular, for vacuole transport and cell wall modulation. As grape berries, and in particular their development, are of great interest for both the research and commercial grapevine communities, we performed a subsequent RNA-Seq experiment on 4 important stages of berry development and 1 stage of post-harvest withering. About 11,000 genes resulted as differentially expressed, 51 of which are specific for Corvina. Our method was effective in detecting putatively important transcripts which could have been missed with reference-based strategies.

TransPLANT, developing a trans-National infrastructure for Plant Genomic Science

Bolser, D.M.[1,*]; **Kersey, P.**[1]

[1]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

*Presenting author: **Daniel M. Bolser** (dbolser@ebi.ac.uk)

Food and energy security are major challenges facing humanity in the coming decades. The falling cost of nucleotide sequencing is creating significant opportunities for crop improvement through plant breeding and increased understanding of plant biology. In particular, progress is being made by interpreting the growing volume of plant genomics data in the context of phenotype. However, at present, there is no adequate infrastructure for plant genomic data.

The transPLANT project aims to develop a new infrastructure for this data, leveraging the experience of medical informatics while addressing the particular challenges and opportunities of plant genomics. Compared with vertebrate genomes, plant genomes may be large and have complex evolutionary histories, making their analysis a hard problem both in terms of theory, and in terms of the resources required for data storage and analysis. Issues include: genome size, polyploidy, and the quantity, diversity and dispersed nature of data in need of integration.

To address these problems, transPLANT will develop distributed solutions, exploiting the expertise of the project partners in particular species and problems, to provide a set of computational and interactive services for the plant research community. These services will be developed on top of the outputs of research activities to building new repositories and developing new algorithms. Input from the plant science and other related communities will be garnered through networking activities and a series of training workshops will educate the community in the use of transPLANT tools and data.

TransPLANT will be built on standard technologies for data exchange and representation, service provision, virtual compute infrastructure, and interface development. Where such standards are currently lacking (as in phenotype description), they will be actively developed in the context of the project.

NGS data processing with Conveyor workflows

Linke, B.[1,*]; Goesmann, A.[1]

[1]Center for Biotechnolgy, Bielefeld University, Bielefeld, Germany

*Presenting author: **Berkhard Linke** (blinke@CeBiTec.Uni-Bielefeld.DE)

Workflows have become a widely used tool in the field of automated data processing and analysis. They allow to split a complex operation into smaller and simpler tasks, while maintaining their logical and temporal dependencies. A number of workflow implementations like Taverna or Galaxy are already available for processing genome and postgenome data.

Conveyor is a novel approach to workflows, with a focus on bioinformatics data analysis. It is based on a statically typed object-oriented data model, allowing inheritance, interfaces, and generic data types. Processing steps are implemented as classes based on that data model, providing a high level of reusability, abstraction and adoption of functionality. A workflow is built from connected instances of processing and input/output steps, ranging from simple pipelines to complex workflow graphs. A multithreaded processing engine allows Conveyor to exploit modern hardware resources and process workflows in a very efficient way.

Plugins allow extending Conveyor with new functionality and thereby adapt it to new data domains. Being a library itself, Conveyor can be easily integrated into other applications, providing them with a flexible and configurable workflow processing engine.

Several ready-to-use plugins exist for the handling of next-generation sequencing data. Reads and similar sequences may be read and written in various formats (Fasta & quality files, FastQ, SFF), stripped, filtered, combined, merged, inserted or modified in any arbitrary way. Creating two distinct processing paths within a single workflow can be used for handling paired end data, allowing e.g. filtering of low quality paired end sequences in interleaved FastQ files. Other plugins handle external applications like Blast searches, read mapping, or SNP and variance calling.

Reference:

Linke, B., R. Giegerich, & A. Goesmann. 2011. "Conveyor: a workflow engine for bioinformatic analyses". *Bioinformatics*, 27(7), 903 - 911.

MotifLab: A tools and data integration workbench for motif discovery and regulatory sequence analysis

Klepper, K.[1]**; Drabløs, F.**[1,*]

[1]Department of Cancer Research and Molecular Medicine,
Norwegian University of Science and Technology (NTNU), Trondheim, Norway

*Presenting author: **Finn Drabløs** (finn.drablos@ntnu.no)

Discovering binding motifs and bindings sites for transcription factors is an important bioinformatics problem in genome research, and many tools have been proposed to search for novel motifs or to scan for potential sites that match established binding motifs. Unfortunately, traditional motif discovery and scanning methods that only rely on sequence data have a tendency to make a lot of false predictions. However, it has been demonstrated that use of additional information, such as gene expression, sequence conservation, location of DNase HS sites and epigenetic marks etc., has the potential to reduce the number of spurious predictions and also discriminate between functional and non-functional binding sites. A lot of data that could prove useful for this purpose is already available at genome-wide scales and more data for different organisms, cell-types and conditions is being published at an increasing rate.

MotifLab is a general software workbench for regulatory sequence analysis designed to make it easy to incorporate different types of data into the motif discovery process. A key application of MotifLab is for constructing positional priors tracks based on various sequence feature annotations. Positional priors can be used to highlight those parts of sequences that are considered more likely to contain functional binding sites, and they can be employed by motif discovery methods to guide the search or be used in a post-processing step to filter unpromising predictions. MotifLab can interface with several popular motif discovery tools (including MEME, Priority, MDscan, Weeder and BioProspector) to predict both individual binding sites and combinations of sites that could potentially function together (cis-regulatory modules).

MotifLab is available from http://www.motiflab.org for download and local installation, or it may be launched via Java web start.

Genome-wide methods for the study of DNA-methylation inheritance

Colomé-Tatché, M.[1,*]; Cortijo, S.[2]; Wadenaar, R.[1]; Jansen, R.C.[1]; Colot, V.[2] and Johannes, F.[1]

[1]Groningen Bioinformatics Centre, Faculty of Mathematics and Natural Sciences, University of Groningen, Nijenborgh 7, 9747 AG Groningen, The Netherlands
[2]Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique (CNRS) UMR8197-Institut National de la Santé et de la Recherche Médicale (INSERM) U1024, 46 rue d'Ulm, 75230 Paris cedex 05, France

*Presenting author: **Maria Colomé-Tatché** (m.colome.tatche@rug.nl)

Inter-individual differences in DNA methylation states can provide a source of heritable phenotypic variation independent of DNA sequence changes [1]. Accumulating evidence suggests that this mode of epigenetic inheritance is more widespread in plant populations than previously appreciated. We have begun to characterize the epigenetic basis of complex trait inheritance in an experimental system of so-called Epigenetic Recombinant Inbred Lines (EpiRILs) of the plant Arabidopsis [2,3,4]. This population was derived from two parents with nearly identical DNA sequences but drastically divergent DNA methylation profiles.

Here I highlight our analytical and computational approach for characterizing the full methylome of individuals in this population using bisulfite sequencing (BS-seq) and whole-genome tiling arrays (MeDIP-chip). I further demonstrate how this information can be used to built genome-wide maps of Differentially Methylated Regions (DMRs) and how these DMRs can faciliate inferences about global recombination patterns in this nearly isogenic population.

Our approach represents a general strategy and can be applied to other populations and scenarios.

[1]F. Johannes, V. Colot and R.C. Jansen, Nature Reviews Genetics 9, 883 (2008).

[2]F. Johannes, E. Porcher, F.K. Teixeira, V. Saliba-Colombani, M. Simon, et al., PLoS Genetics 5(6): e1000530. doi:10.1371/journal.pgen.1000530 (2009).

[3]F. Roux, M. Colomé-Tatché, C. Edelist, R. Wardenaar, P. Guerche, F. Hospital, V. Colot, R.C. Jansen and F. Johannes, Genetics, 188:1015-1017, doi:10.1534/genetics.111.128744 (2011).

[4]F. Johannes and M. Colomé-Tatché, Nature Reviews Genetics 12, 376 (2011).

A statistical approach to estimate copy number from NGS capture sequencing data

Rigaill, G.[1,*]; Kluin, RJC[2]; Xue, Z.[3]; Bernards, R.[3]; Majewski, IJ[3]; Wessels, LFA[1]

[1]Bioinformatics and Statistics and Cancer Systems Biology Center (CSBC),
The Netherlands Cancer Institute, Amsterdam, The Netherlands
[2]Central Microarray Facility, The Netherlands Cancer Institute, Amsterdam, The Netherlands
[3]Department of Molecular Carcinogenesis, The Netherlands Cancer Institute,
Amsterdam, The Netherlands

*Presenting author: **Guillem Rigaill** (g.rigaill@nki.nl)

Target enrichment, also referred to as DNA capture, provides an effective way to focus sequencing efforts on a genomic region of interest. Capture data is typically used to detect single nucleotide variants. It can also be used to detect copy number alterations (CNAs), which is particularly useful in the context of cancer, where such changes occur frequently.
In copy number analysis of capture sequencing data it is common practice to determine logratios between test and control samples, but this approach results in a loss of information as it disregards the total coverage at a locus.

We instead modeled the coverage of the test sample as a linear function of the control samples. This approach is able to deal with regions that are completely deleted, which are problematic for methods that use log ratios. To demonstrate the utility of our approach, we used capture data to determine copy number for a set of ~600 genes in a panel of nine breast cancer cell lines. We found high concordance between our results and those generated using a SNP genotyping platform. When we compared our results to other methods, including ExomeCNV, we found that our approach produced better overall correlation with SNP data.

Variability in RNA-seq: design and modeling

McIntyre, L.M.[1,*]

[1]Molecular Genetics and Microbiology, University of Florida

*Presenting author: **Lauren M. McIntyre** (mcintyre@ufl.edu)

RNA-seq is revolutionizing the way we study transcriptomes. Alternative splicing of transcript isoforms and allele specific expression are two applications of this new technology that are particularly exciting. Reports of differences in exon usage, and splicing between samples as well as differences among alleles and the best way to model and quantify these differences are a subject of great interest. This new technology has novel challenges. Some challenges are bioinformatics (e.g. map bias); some are technical (e.g. lane to lane variability), and some are statistical (e.g. appropriate models for allele specific expression). Without proper consideration of all three sets of challenges bioinformatic, technical and statistical, inferences may be misleading. This presentation focuses on steps for dealing with these challenges that reduce or eliminate some of these concerns. Statistical models for allele specific expression and alternative splicing are presented.

A Comprehensive Evaluation of Normalization Methods for High-Throughput RNA Sequencing Data Analysis

Aubert, J.[1,2,*]; Dillies, MA[3]; Rau, A.[4]; Hennequet-Antier, C.[5] ; Jeanmougin, M.[6,7]; Servant, N.[8,9,10]; Keime, C.[11] ; Le Crom, S.[12,13,14] ; Guedj, M.[6] and Jaffrézic, F.[4], on behalf of the French StatOmique Consortium

[1]AgroParisTech, UMR518 Mathématiques et Informatique Appliquées, Paris, F-75005 France
[2]INRA, UMR518 Mathématiques et Informatique Appliquées, Paris, F-75005 France
[3] Institut Pasteur, PF2 – Plate-forme Transcriptome et Epigénome, Paris, F-75724 France
[4] INRA, UMR1313 Génétique Animale et Biologie Intégrative, Jouy-en-Josas, F-78352 France
[5]INRA, UR83 Recherches Avicolesn Nouzilly, F-37380 France
[6]Pharnext, Department of Biostatistics, Issy-les-Moulineaux, F-92130 France
[7]Laboratoire Statistique et Génome, Université d'Evry Val d'Essonne, UMR CNRS 8071 – USC France
[8]Institut Curie, Paris, F-75248 France
[9]INSERM, U900, Paris, F-75248 France
[10]Ecole des Mines de Paris, Fontainebleau, F-77300 France
[11]Institut de Génétique et de Biologie Moléculaire Cellulaire, CNRS UMR 7104, INSERM U596, Université de Strasbourg, Ilkirch, France
[12]Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, F-75005 France
[13]INSERM, U1024, Paris, F-75005 France
[14]CNRS, UMR8197, Paris, F-75005 France

*Presenting author: **Julie Aubert** (julie.aubert@agroparistech.fr)

Background

During the last three years, a number of approaches for the normalization of RNA sequencing data have emerged in the literature, differing both in the type of bias adjustment and in the statistical strategy adopted. However, as data continue to accumulate, there has been no clear consensus on the appropriate normalization method to be used or the impact of a chosen method on the downstream analysis.

Results

In this work, we focus on a comprehensive comparison of seven recently proposed normalization methods for the differential analysis of RNA-seq data, with an emphasis on the use of varied real and simulated datasets involving different species and experimental designs to represent data characteristics commonly observed in practice.

Conclusion

Based on this comparison study, we propose practical recommendations on the appropriate normalization method to be used and its impact on the differential analysis of RNA-seq data.

Model-based clustering for high-throughput sequencing data to determine similar expression profiles across genes

Papastamoulis, P.[1]; Rau, A.[2]; Maugis-Rabusseau, C.[3]; Celeux, C.[4]; Martin-Magniette, M.L.[1,5] *

[1]UMR INRA 1165/UEVE, ERL CNRS 8196, Unit of Plant Genomics, Evry, France
[2]INRA, UMR 1313 GABI, Jouy-en-Josas, France
[3]Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, France
[4]Inria Saclay - Île-de-France, Orsay, France
[5]UMR AgroParisTech/INRA 518 Mathématiques et informatiques appliquées, Paris, France

*Presenting author: **Marie Laure Martin-Magniette** (marie_laure.martin@agroparistech.fr)

Background

In recent years gene expression studies have increasingly made use of next generation sequencing technology, and in turn, research concerning the appropriate statistical methods for the analysis of digital gene expression (DGE) has flourished. In this work, we focus on the question of clustering DGE measures (i) to discover groups of co-expressed genes with similar expression profiles (ii) to identify groups of genes displaying similar behavior with respect to reference samples. Clustering analyses based on metric criteria such as the K-means algorithm and hierarchical clustering have been used in the past to cluster microarray-based measures of gene expression as they are rapid, simple, and stable. However, such methods require both the choice of metric and criterion to be optimized, as well as the selection of the number of clusters. An alternative to such methods are probabilistic clustering models.

Results

For the first task, we propose a mixture of Poisson loglinear models to cluster count-based DGE observations. A set of simulation studies compares the performance of the proposed model with that of two previously proposed approaches for Serial Analysis of Gene Expression data. For the second task, we propose a mixture of Poisson Generalized Linear Models to find groups of genes behaving similarly with respect to a set of reference samples. This latter model has also the potential to incorporate additional biological information (e.g., the gene's length). For a given number of mixture components, parameters in both models are estimated using the EM algorithm and the optimal number of clusters is chosen according to the Integrated Completed Likelihood (ICL) criterion. R packages are available for both models and their performances are illustrated on real high-throughput sequencing data.

Conclusion

Model-based clustering provides a rigorous framework to perform gene clustering. It has the advantage of providing straightforward procedures for parameter estimation as well as an a posteriori probability for each gene of belonging to each cluster. Moreover model selection criteria are available to choose the number of groups without a priori knowledge.

Haplotype estimation and imputation using low-coverage sequence data

Marchini, J.L.[1,*]; Delaneau, O.[1,2]; Zagury, JF.[2]

[1]Department of Statistics, University of Oxford, UK
[2]Chaire de Bioinformatique, Conservatoire National des Arts et Metiers, Paris, France

*Presenting author: **Jonathan L. Marchini** (marchini@stats.ox.ac.uk)

The determination of SNP genotypes from low-coverage sequencingdata is a challenging statistical problem. The task can be described as an inverse problem where we indirectly observe a set of genotypes via sequencing in a sample of individuals and wish to convert these genotypes into the underlying (unobserved) haplotypes carried by the study samples. Estimation of the haplotypes determines the set of genotypes carried by each individual in the sample. Accurate models and methods and needed to infer the haplotype and genotypes and we have recently developed a new method called SHAPEIT [1] which has improved accuracy and computational efficiency compared to several other methods. Some notable features of the method are that it (a) scales linearly in the number of haplotypes used in the update step at each iteration and in the number of SNPs and number of samples, (b) can phase whole chromosomes at once, and (c) can handle mother-father-child trios and parent-child duos. We will describe the details of this algorithm and also a new version SHAPEIT2, that fuses that haplotype subset selection ideas in IMPUTE2 [2], to further boost performance and speed. We illustrate this performance using six different large sample, whole chromosome datasets.

[1] O. Delaneau, J. Marchini, JF. Zagury (2011) A linear complexity phasing method for thousands of genomes. Nature Methods doi:10.1038/nmeth.1785.

[2] B. Howie, P. Donnelly, J. Marchini (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genetics 5(6): e1000529 .

SNP discovery from Next Generation Sequencing to study for genome variation in Arabis alpina natural populations

Odong, T.L.[1,*]; Kourmpetis, Y.[1]; Wunder, J.[2]; Coupland, G.[2]; Per, T.[3]; Ågren, J.[3]; Herzog, M.[4]; Obeso, JR[5]; Bink, M.C.A.M.[1]

[1]Wageningen University and Research, Lab of Bioinformatics & Biometris, Wageningen, The Netherlands

[2]Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany

[3]Evolutionary Biology Center, University of Uppsala, Uppsala, Sweden

[4]Centre National de la Recherche Scientifique, CNRSUniversité Joseph FourierLaboratoire d'Ecologie Al-pine, LECA, Grenoble, France

[5]University of Oviedo, Ecology Unit, Oviedo, Spain

*Presenting author: **Thomas Lapaka Odong** (thomas.odong@wur.nl)

We used Next Generation Sequencing (NGS) technology for SNP discovery to study genomic variation in the Arabis alpina. Arabis alpina is a perennial plant species of great interest in genetics and molecular biology. It is closely related to Arabidopsis thaliana, a well-studied annual plant, and as such Arabis alpina has a potential for becoming a model plant for studying molecular and genetic mechanisms of perennial plants. The overall aim of this project is to unravel the molecular and genetic basis of the perennial life strategies of Arabis alpina. For our study, eight natural populations were selected from different geographic locations across Europe (Sweden(5), Spain(1) and France(2)). In this talk we present the result of the analysis of NGS and the challenges encountered during the analysis.

The genomes of 22 individuals, across the eight European populations were sequenced by Next Generation Sequencing (NGS) technology in two batches of 16 and 6 individuals respectively. The second batch comprised only Swedish individuals. The raw data (reads) obtained from this project were mapped using BWA to the reference genome of Arabis alpina, which was assembled (yet unpublished) at the Max Planck Institute in Cologne, Germany. We used SAMtools for discovery of Single Nucleotide Polymorphisms (SNPs). We discovered about 3 million SNP distributed across the genome of Arabis alpina. Exploratory analysis showed that the SNPs discovered provide a good general description of spatial population structure of the 22 individuals. However, the genetic distances within the Swedish subpopulations were inconsistent, hinting to batch effect in our sequence data. An in depth evaluation of the different steps in the analysis, e.g., thresholds on read depths, might reveal plausible causes of the unexpected Swedish population structure. Furthermore, additional NGS data has recently been generated on other individuals from the same populations. These data will be helpful to disentangle the problems with the initial NGS data and to arrive at a large set of reliable SNPs. In the next steps, we will embark on the selection of small subset of SNPs that will be used for other objectives (studying population structure, outcrossing/clonal rate) of the project.

Genotyping by sequencing tetraploid potato and attempts towards reconstruction of haplotypes

de Boer, J.[1,*]; Uitdewilligen, J.[1,3]; Eilers, P.[4,5]; Wolters, AM[1,3]; Paulo, J.[2,4]; Vos, P.[1,2]; Visser, R.[1,2,3]; van Eeuwijk, F.[4]; van Eck, H.J.[1,2,3]

[1]Wageningen University laboratory of Plant Breeding, Wageningen, The Netherlands
[2]Centre for BioSystems Genomics, Wageningen, The Netherlands
[3]The Graduate School for Experimental Plant Sciences, Wageningen, The Netherlands
[4]Biometris, Wageningen University, Wageningen, The Netherlands
[5]Erasmus University Medical Center, Rotterdam, The Netherlands

*Presenting author: **Jan de Boer** (janmdeboer@gmail.com)

The genotyping of DNA sequence variants in heterozygous polyploid species such as potato, is more challenging than in diploid species because a given gene may be represented by a number of different alleles with different zygosity (nulliplex, simplex, duplex, triplex, quadruplex). Data were generated with the Illumina HISeq2000, 100bp paired-end reads. Differences in relative read depth of alleles was used to estimate the allele copy-number. Unphased SNPs allowed genome wide association (GWAS) analysis and identified QTL for various traits.

Our ambition is to improve the detection of marker trait associations. Therefore we aim to identify each of the haplotypes at QTL loci and to estimate the magnitude of their effect on various traits. Unfortunately tetraploids will not easily disclose haplotype information at a specific genetic locus. Haplotypes can be obtained (1) statistically with haplotype probability estimation, and/or (2) with bioinformatics e.g. read-backed approaches and combinations thereof. We will discuss that GWAS with haplotypes will recover some of the "missing heritability" which is not explained by unphased SNPs.

This research is supported by grants of the Centre for Biosystems Genomics and the Dutch Technology Foundation STW: WPB-7926.

Genotyping by Sequencing in Maize

Glaubitz, J.C.[1,*]; Bradbury, PJ[2]; Harriman, J.[1]; Sun, Q.[3]; Casstevens, TM[1]; Elshire, RJ[1]; Acharya, CB[1]; Mitchell, SE[1]; Zhang, Z.[1]; Romay, MC[1]; Buckler, ES[1,2]

[1]Institute for Genomic Diversity, Cornell University, Ithaca NY 14853 USA
[2]United States Department of Agriculture, Agriculture Research Service, Ithaca NY 14853 USA
[3]Computational Biology Service Unit, Cornell University, Ithaca NY 14853 USA

*Presenting author: **Jeff Glaubitz** (jcg233@cornell.edu)

Genotyping by sequencing (GBS) is a reduced representation, next generation sequencing approach that provides a robust and cost-effective means to genotype large numbers of individuals at high density, by targeting sequence adjacent to restriction enzyme cut sites. The robustness and cost-effectiveness of the GBS protocol derives from its relative simplicity. To date, we have genotyped more than 20,000 maize samples via GBS. We have built our own bioinformatics pipeline for the analysis of this large data set, a key feature of which is its scalability. The main disadvantage of GBS is the large amount of missing data. To call heterozygous segments, correct errors, and facilitate imputation in biparental populations of recombinant inbred lines (RILs), we have implemented a hidden Markov model (HMM). In contrast, for imputation of unrelated inbred lines, we instead employ a nearest neighbor, fixed window approach. The limited pool of founders contributing to elite US inbreds facilitates their accurate imputation. Furthermore, the bottleneck experienced by maize during its domestication also placed a limit on the number of founding haplotypes. Strategies for improving our imputation percentage and accuracy will be discussed, as will our current results using GBS data for GWAS, NAM-GWAS, and for improving the maize reference genome.

The effect of LD between SNPs on some statistical methods for GWAS

Waldmann, P.[1,2,*]**; Sölkner, J.**[2]

[1]Thetastats, Uardavägen 91, 224 71 Lund, Sweden
[2]Division of Livestock Sciences, Department of Sustainable Agricultural Systems, University of Natural Resources and Applied Life Sciences, Gregor Mendel Str.33, A-1180, Vienna, Austria

*Presenting author: **Patrik Waldmann** (Patrik.Waldmann@boku.ac.at)

The number of publications performing genome-wide association studies (GWAS) has increased dramatically. The large number of SNPs (p) in contrast to the relatively small number of individuals (n) introduces a statistical problem often referred to as the to the $n \ll p$ problem. Penalized multiple regression and Bayesian variable selection approaches have been developed to overcome the challenges introduced by this problem. Recently, it has been shown that some of these methods are sensitive to correlated predictors. Based on simulation studies, we show here that two popular approaches, the lasso and stochastic search variable selection (SSVS), breaks down when some SNPs are in moderate to high linkage disequilibrium (LD) with each other. Hence, methods that account for LD, for example the elastic net, should be used in GWAS.

Meta-statistics for Biomarker Selection in the Omics Sciences

Wehrens, R.[1,*]; Franceschi, P.[1]

[1]Biostatistics and Data Management, Fondazione Edmund Mach

*Presenting author: **Ron Wehrens** (ron.wehrens@fmach.it)

Background
Biomarker selection, i.e., the definition of which variables are important in statistical regression or discrimination models, is an ever more important topic in the omics sciences. Data from these fields are typically characterized by a low number of samples, but a large number of variables – a meaningful biological interpretation often is only possible when considering the most important variables.

Methods
In this context, statistical tests like the t test will lead to many false positives, while multiple testing corrections tend to lose much power and select only very few variables. In addition, the cutoff value (usually set to a value like 5%) is often chosen in a haphazard way. We present two meta-statistics to tackle the problem of variable selection: higher criticism thresholding [1,2] and stability selection [3,4]. Higher criticism thresholding, applicable in a two-class discrimination setting, is a way to set suitable cutoff levels for significance, based on the data at hand. The underlying mechanism has been described as the "z-score of the p-value" [1]. The current work has extended higher criticism to multivariate methods like PLSDA and the VIP statistics [4]. Stability selection is a novel variable selection method, assessing the stability of biomarker selections under perturbations of the data. The concept is extremely general and robust and can be applied both in regression and discrimination cases: primary selection methods assessed in thie work include PLS and lasso models.

Results
Simulated as well as experimental data show very good results for both stability selection and higher criticism. The experimental data in this study consist of LC-MS metabolomics data of spiked-in apple extracts [5] – such spike-in data are extremely important in assessing the value of biomarker selection methods but are rarely available. Good results are also obtained in other areas of science [1-3]. The advantages of stability selection include a broad applicability (regression, discrimination) and modest computational demands; on the other hand, the number of samples that is required is relatively high. For discrimination problems with fewer than, say, eight samples per class, it is probably better to rely on the higher criticism approach. Both higher criticism and stability selection have been implemented in an R package, BioMark, available from the CRAN repository, and also containing the experimental spike-in data.

[1] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.

[2] Donoho, D. and Jin, J. (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *PNAS*, **105**, 14790–14795.

[3] Meinshausen, N. and Buhlmann, P. (2010). Stability selection. *J. R. Statist. Soc. B*, **72**, 417–473. With discussion.

[4] Wehrens, R., Franceschi, P., Vrhovsek, U., and Mattivi, F. (2011). Stability-based biomarker selection. *Anal. Chim. Acta*, **705**, 15–23.

[5] Franceschi, P., Masuero, D., Vrhovsek, U., Mattivi, F., and Wehrens, R. (2012). A benchmark spike-in data set for biomarker identification in metabolomics. *J. Chemom.*, **26**, 16-24.

Sequential Monte Carlo algorithms for high throughput genetic mapping from dense genomewide SNPs

Kruijer, W.T.[1,*]; Schafer, C.[2,3]; Malosetti, M.[1]; ter Braak, C.J.F[1];
van Eeuwijk, F.E.[1]; Bink, M.C.A.M.[1]

[1]Biometris, Wageningen University and research Centre, the Netherlands
[2]Université Paris Dauphine, France
[3]Centre de Recherche en Économie et Statistique, Malakoff, France

*Presenting author: **Willem Kruijer** (willem.kruijer@wur.nl)

A major outcome of Next Generation Sequencing are the abundant availability of SNP markers in humans and many agricultural species. These amounts of SNPs pose new challenges for the statistical tools for genetic mapping of complex traits.

Bayesian variable selection (BVS) methods are well established in quantitative genetics. Considering a linear model for a quantitative trait and a large number of genetic predictors, the quantity of interest is the posterior mean on the binary model space (probability that a marker is included or not). In genomic selection BVS is used for prediction, whereas in other contexts the goal is variable selection as such.

Despite various successful applications, most BVS methods suffer from two important drawbacks. First, they typically rely on MCMC algorithms with local transitions, which do not easily allow to profit from parallel computing environments. Secondly, Markov MCMC algorithms may converge extremely slow and produce poor estimates of the actual posterior mean. This becomes very predominant when the predictors are highly correlated (e.g. high marker density) and even more when epistatic interactions are included in the model.

Although these problems are somewhat alleviated by recent MCMC-improvements (Bottolo and Richardson, 2010) we advocate Sequential Monte Carlo (SMC) algorithms (Del Moral et al. 2006). These were adapted to BVS by Schäfer and Chopin (2011). They constructed global, fast-mixing adaptive transition kernels with independent proposals, drawn from a suitable parametric family. This allows for reliable sampling from highly multi-modal distributions and for massive parallelisation.

Using SMC we analyse real and simulated data from natural- and experimental populations, such as flowering time measured in the arabidopsis hapmap-collection, genotyped at 214553 SNP markers. For segments of several thousands of markers, we compared SMC with other MCMC-based implementations of BVS (such as Rqtl/bim), and with non-Bayesian methods such as composite interval mapping. We found that for strongly correlated markers SMC produces much more accurate estimates of QTL-locations. Depending on the underlying genetic architecture it also requires fewer evaluations of the target posterior. We expect that SMC can handle larger problems too; in the near future this will be tested on a larger cluster.

Under the hood of the 1000 Genomes Project

DePristo, M.A.[1,*]

[1]The Broad Institute of MIT and Harvard

*Presenting author: **Mark A. DePristo** (depristo@broadinstitute.org)

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. By combining low coverage whole genome sequencing, targeted exome capture, and array genotype data on over 1000 individuals with a worldwide geographic distribution, we have generated an integrated map of over 40 million SNPs, short indels and larger structural variants in the human genome. I discuss the evolution of the next-generation sequencing (NGS) data underlying the public 1000 Genomes Project resource, from some of the earliest technologies of 2009 to today's state-of-the-art data from Illumina, Life Technologies, and Pacific Biosciences. I highlight key NGS analytic advances originating from the Project, including the BAM and VCF file formats, multi-sample variation discovery and genotyping algorithms, large-scale validation assays using multiple independent technologies, and the integration of disparate variant types, from SNPs to structural variants, using a unified likelihood-based imputation framework. Finally I outline lessons learned from the Pilot and Phase I freezes and outstanding challenges in variant calling and integration for population-scale sequencing that are driving the future directions of the 1000 Genomes Project.

Towards Arabidopsis thaliana genetic regulatory network using discrete Bayesian network

Vandel, J.[1,*]; De Givry, S.[1]; Leroux, D.[1]; Vignes, M.[1]; Loudet, O.[2]; Martin-Magniette, M.L.[3,4]; Mangin, B.[1]

[1]INRA, UR 875, Unité de Biométrie et Intelligence Artificielle, F-31326, Castanet-Tolosan, France
[2]INRA, UMR 1318, Institut Jean-Pierre Bourgin, F-78000 Versailles, France
[3]INRA, UMR 1165, Unité de Recherche en Génomique Végétale, F-91057, Evry, France
[4]INRA, UMR 518, Mathématiques et Informatique Appliquées, F-75231 Paris, France

*Presenting author: **Jimmy Vandel** (jimmy.vandel@toulouse.inra.fr)

Modern genomic technologies are giving access to genotype and genomic data measured on the same sample. Combining these different observations to discover regulations or interactions is the challenge of the genetical genomics.

Using both genetic and genomic data sets observed on *Arabidopsis thaliana* Recombinant Inbred Lines (RIL), we predicted a gene regulatory network using the framework of discrete static Bayesian networks. The algorithm was based on a hill climbing greedy search on the space of directed acyclic graphs and it used the Bayesian uniform Dirichlet score.

The 158 RILs were genotyped for 89 SNP markers regularly spaced on the five *Arabidopsis* chromosomes. We inferred, using the R package eqtl, pseudo-markers spaced every 1 cM that were discretized in three classes. 34660 transcript levels were observed using the CATMA microarrays. We used to predict the network, 4176 expression levels those that were genetically explained by at least one significant eQTL. Missing data were completed and expression levels were discretized in at most four classes.

The predicted network was then compared to the eQTL analysis in term of cis and trans gene regulation. 68% of surely *cis* regulated genes were linked to a *cis* (pseudo) marker by an oriented path in the predicted network. A nearly equal rate (69%) was found for trans regulated genes. Moreover, we found some predicted interactions that are known as direct interactions or that had common GO terms.

Genomic Breeding using variomics data and decision support systems

Buntjer, J.[1,*]

[1]Keygene N.V., Wageningen, The Netherlands

*Presenting author: **Jaap Buntjer** (jaap.buntjer@keygene.com)

With the rapidly increasing performance of sequencing and genotyping technology, we are on the dawn of an era in which access to data that represent the genomewide genotypic variation will become reality for many economical important crops.
On the one hand, this will lead to the identification of many causal genomic factors that contribute to economically important traits. On the other hand, we will get control over highly complex traits by selection using an abundance of markers, without the need for knowledge on the causal biological mechanisms.

In both ways, for the plant breeding practice this means that the actual breeding process itself is going to be the limiting factor, rather than the creation of knowledge and molecular data. In the current presentation, an alternative breeding strategy to both MAS and GWS is proposed, based on deterministic genotype predictions.

POSTERS

A framework for characterization of cultivar-specific transcriptome from next generation sequencing data

Abate, F.[1]; Giugno, R.[2]; Bombieri, N.[3]; Ferrarini, A.[4]; Ficarra, E.[1]; Pulvirenti, A.[2]; Delledonne, M.[4]; Acquaviva, A.[1,*]

[1]Politecnico di Torino, Department of Control and Computer Engineering, Torino, Italy
[2]Università degli Studi di Catania, Department of Clinical and Molecular Biomedicine, Catania, Italy
[3]Università degli Studi di Verona, Department of Computer Science, Verona, Italy
[4]Università degli Studi di Verona, Department of Biotechnology, Verona, Italy

*Presenting author: **Andrea Acquaviva** (andrea.acquaviva@polito.it)

Genetic variation is at the base of phenotypic variation among different varieties of the same plant species. Until recently genetic differences have been described in the context of the reference genome. However recent findings allowed by the extensive use of next generation sequencing (NGS) have shown that different ecotypes and varieties can possess a large number of highly divergent loci and genes which are specific of a single variety or which are not in common with the rest of the species. Same situation has been highlighted in humans in which for example population specific sequences have been identified by the sequencing of African and Asian genomes. All this findings strongly suggest that genetics and transcriptomics (analysis of alternative splicing) must be performed in the context of individual genomes. However a full genome assembly still present problems due to highly repetitive sequences which cannot be easily solved with current technologies. Therefore we propose, an innovative computational infrastructure, based on cloud computing and efficient intra-node acceleration, for assembly of transcriptomes and direct comparison with reference genome and among different assembled transcriptomes to allow the characterization of differences in coding sequences, alternative splicing, etc., among different ecotypes and varieties of plants. This infrastructure will either push a step forward the efficiency of cloud computing applied to NGS data by introducing new features, and will propose hardware accelerators for cloud nodes. On the algorithmic viewpoint, innovative approaches will be proposed to carry out the comparison of reconstructed transcriptomes with reference genome and related annotations. Annotations will be carried out bringing together in an unique suite all advanced techniques enriched with novel data mining and probabilistic approaches managing information from Gene Ontology, RNA Interference, Transcription Factors, Pathways and Biological networks.

RNA-Seq data normalization methods for differential gene expression analysis

Maza, E.[1,2,*]; Frasse, P.[1,2]; Senin, P.[1,2] ; Fu, Y.[1,2]; Bouzayen, M.[1,2]; Zouine, M.[1,2]

[1]Université de Toulouse, INP-ENSA Toulouse, Génomique et Biotechnologie des Fruits,
Avenue de l'Agrobiopole BP 32607, Castanet-Tolosan F-31326, France
[2]INRA, Génomique et Biotechnologie des Fruits, Chemin de Borde Rouge, Castanet-Tolosan,
F-31326, France

*Presenting author: **Elie Maza** (Elie.Maza@ensat.fr)

Background

It is now a commonplace to consider that RNA-Seq data (from high throughput sequencing technologies) need to be normalized prior to any quantitative analysis required for the extraction of differentially expressed genes. The present study is dealing with RNA-Seq data set aiming at profiling the fruit set process in the tomato. RNA was isolated at three stages (flower buds, anthesis and post-anthesis) and three biological replicates were subjected to Illumina mRNA-Seq technology sequencing. The generated data set is normalized using the most widely used normalizations methods (Total Counts, FPKM, Upper Quartile, Median, TMM and RLE) with software R and dedicated packages (edgeR and DESeq). The objective is to compare the output of each of these methods. Moreover, we also propose a new normalization method called Mode Ratio Normalization based on its better suitability to a wide variety of differential expression situations.

Results

Differential expression analyses, carried out following application of the normalization methods mentioned above, lead to up to 50% of non common differentially expressed genes. We show also that the ratios obtained are different between the different normalization methods.
We then studied the impact of the normalization methods on simulated data sets to achieve more accurate comparisons of the given methods.

Conclusions

All methods seem obviously to be well suited for detection of high differentially expressed genes. However, the importance of the normalization method is greater for less important differential expression levels with ratios ranging between 1 to 4. The comparative use of different normalization methods to process simulated data sets reveals that the Mode Ratio Normalization is the most accurate method.

Evaluation of transcriptomic data and identification of putative stress responsive genes

Janská, A.[1,2]; Aprile, A.[3]; Zámečník, J.[1]; Cattivelli, L.[4]; Ovesná J.[1,*]

[1]Crop Research Institute, v.v.i., Drnovská 507, 161 06 Prague 6, Czech Republic
[2]Charles University in Prague, Faculty of Science, Viničná 5, 128 44 Prague 2, Czech Republic
[3]University of Salento, Department of Biological and Environmental Sciences and Technologies, Ecoteckne prov.le Monteroni-Lecce, 73100 Lecce, Italy
[4]Agricultural Research Council of Italy, Genomics Research Centre, via S. Protaso, 302, I -29017 Fiorenzuola d'Arda PC, Italy

*Presenting author: **Jaroslava Ovesná** (ovesna@vurv.cz)

Background

We report a series of microarray-based comparisons of gene expression in the leaf and crown of the barley cultivars Luxor, Igri and Atlas 68 following the exposure of young plants to various periods of low (above and below zero) temperatures. A transcriptomic analysis identified genes which were either expressed in both the leaf and crown, or specifically in one or the other. Because array data analysis is very complex, we report several algorithms we deciced to use.

Results

Data from Affymetrix DNA arrays were evaluated in three biological replicates of each sample. Data were statistically evaluated by RMA (Robust Multi-array Average) (Irizarry et al. 2003) with the use of "R package Affymetrix library" (Irizarry et al. 2006), MAS 5.0 algoritm and Genespring GX 7.3 software. Expression changes (fold change higher than 2) were then evaluated by the MapMan software, which enables to see transcriptomic changes in the network of metabolic pathways ((http://mapman.gabipd.org; Thimm et al. 2004; Usadel et al. 2005). Results were correlated with GeneSpring outputs. DNA array contains more sequences for one target gene, but some of them showed different expression pattern, which suggest presence of multiple gene isoforms. Therefore sequences which generated similar expression profile were called as fully redundant, while sequences with different expression profiles were evaluated as isoforms.

Conclusions

Differences in expression pattern between the crown and leaf were frequent for genes involved in certain pathways responsible for osmolyte production, e.g. sucrose and starch, raffinose, γ-aminobutyric acid metabolism, but also for genes involved in sugar signalling (trehalose metabolism) and secondary metabolism (lignin synthesis).

Acknowledgment

This research was funded by the Czech Republic National Agency for Agricultural Research (project no. QH 81287), by the Czech Ministry of Agriculture (project no. Mze0002700604) and by the Italian Ministry of Agriculture, Special Project ESPLORA.

RNA-Seq analysis of grapevine induced resistance

**Perazzolli, M.[1],*; Moretto, M.[1]; Fontana, P.[1]; Ferrarini, A.[2]; Velasco, R.[1];
Delledonne, M.[2]; Pertot, I.[1]**

[1]IASMA Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1,
38010 San Michele all'Adige (TN), Italy
[2]Università degli Studi di Verona, Dipartimento di Biotecnologie, Strada Le Grazie 15,
37134 Verona, Italy

*Presenting author: **Michele Perazzolli** (michele.perazzolli@fmach.it)

Induced systemic resistance (ISR) is a mechanism of the plant immune system. ISR is activated by selected strains of non-pathogenic microorganisms and provide protection against different types of pathogens in several plant species. In grapevine, treatment with the biocontrol agent *Trichoderma harzianum* T39 (T39) induces resistance against downy mildew caused by Plasmopara viticola. ISR seems to be a promising strategy for controlling crop diseases, but scarce information is available on the molecular mechanisms in non-model plants.

Transcriptional changes associated with T39 treatment and subsequent inoculation with *P. viticola* were analyzed in grapevine by Illumina RNA-Seq method. Three biological replicates were analyzed for each condition. Each biological replicate was sequenced twice on separate lane and paired-end reads 100 nucleotides in length were obtained. More than 15 million reads were obtained for each biological replicate, corresponding to a coverage of at least 32x the grapevine transcriptome. Filtered reads were mapped to the grapevine genome using TopHat tool, and the expression value of grapevine genes was calculated using Cufflinks tool. Whereas exons comprise the 9% of the genome, 77% of mapped reads showed matches to predicted genes. From one to nine isoforms were recognized for each gene, and more than 3500 new expressed regions were identified. Pearson correlations were greater than 0.97 and 0.95 between technical and biological replicates, respectively. Counts of technical replicates were summed to get better coverage, and 7024 genes resulted as differentially expressed in at least one comparison accordingly to DESeq statistical analysis. Functional annotation of differentially expressed genes by Argot2 tool highlighted a specific transcriptional reprogramming of T39-treated grapevines in response to pathogen inoculation.

Transcriptomic profiling of tomato fruit set by deep sequencing uncovers de novo genes

Zouine, M.[1,2,*]; Frasse, P.[1,2]; Maza, E.[1,2]; Senin, P.[1,2]; Fu, Y.[1,2] and Bouzayen, M.[1,2]

[1]Université de Toulouse, INP-ENSA Toulouse, Génomique et Biotechnologie des Fruits,
Avenue de l'Agrobiopole BP 32607, Castanet-Tolosan F-31326, France
[2]INRA, Génomique et Biotechnologie des Fruits, Chemin de Borde Rouge, Castanet-Tolosan,
F-31326, France

*Presenting author: **Mohamed Zouine** (mohamed.zouine@ensat.fr)

Background

Tomato fruit set is a key process with a great economic impact on crop production. We used RNA-seq to investigate the transcriptome dynamics of fruit initiation. RNA were isolated from flowers at bud, anthesis and post-anthesis stages. cDNA libraries were generated from three biological replicates and subjected to Illumina mRNA-Seq technology sequencing.

Results and conclusions

We generated a total of 287.5 millions of 101-bp paired-end high quality reads: 94,91 and 102 Mreads from Bud, Anthesis and Post-Anthesis, respectively. Mapping of these reads to the Tomato genomic sequence was performed by the TopHat software [citation]. More than 90% of the reads were aligned to the genomic sequence.

By successively aligning reads coming from all nine samples, we estimated the cumulative genome and the annotated exome coverage. We showed that when the number of reads increased, the genome and the annotated exome coverage curves approached to a plato, suggesting a sufficient depth of coverage. From the piling up of short reads mapped on the genome, Cufflinks comprehensively predicted 26,376 transcripts, including 21,735 annotated and 4,641 unannotated transcripts in the current tomato annotation database iTAG2.30. Differentially expressed genes were extracted with DegSeq using a multiseries time course design. 3090 genes showed changes throughout Bud to flower transition and 4800 throughout flower to fruit set. Regarding that plant hormones play a crucial role in fruit initiation, we investigate in more details the expression profile of the genes that are known to be involved in hormones biosynthesis and response. We showed that auxin and ethylene related genes are the more represented among this class of genes suggesting that these two hormones play a predominant role in fruit set.

PyRwise: A database and analysis system for RNA and Protein expression data from diverse technologies

Schurch, N.J.[1,*]; Schofield, P.G.[1]; Barton, G.J.[1]

[1]University of Dundee, Dundee, UK

*Presenting author: **Nick Schurch** (nschurch@dundee.ac.uk)

Background
One of the most commonly asked questions in biological research is; 'What changes when I perturb this system?'. The system might be a whole organism, tissue or cell line, while the perturbation could be due to mutation, siRNA knock-down, drug-treatment, stress (e.g. heat shock) or a host of other stimuli. The read-out can be at the level of RNA by microarray, Direct RNA Sequencing (DRS e.g. Helicos Bio), indirect sequencing (e.g. Illumina) or at the level of protein by quantitative mass-spectrometry. Many commercial and open-source tools exist with a focus on one technique, but increasingly, multiple high-throughput experiments from multiple techniques are combined to probe a single biological question. Accordingly, there is a need for software that can provide effective ways to store, organize and analyse comparative data from multiple experiments and technologies.

Results
We present PyRwise; a system designed to assist both wet-lab biologists, and bioinformaticians, in asking questions of differential expression data, both in individual experiments and across multiple experiments. At its core is a database with details of each experiment, raw data, pre-calculated differential expression data, and common annotations. The PyRwise web-portal provides an interface to these data; enabling swift examination and filtering of differential expression results. The data are richly annotated with feature information from Ensembl and function annotations from GO (Gene Ontology).
PyRwise is largely agnostic to the source of the data and can analyse any data that measures differential expression, from microarrays to proteomics and beyond. Although the web-portal provides some simple exploration tools the underlying database is directly scriptable, making it considerably more straightforward to ask complex questions of the data, particularly across a wide range of experiments and a wide range of technologies. Best of all PyRwise is an open-source project written primarily in python and R, and is extensible and modifiable.

Conclusion
Although still in development, PyRwise has already been used to interrogate data from a series of chick embryo development experiments that spans multiple platforms including DRS data and microarray data, and to contrast the results of commonly used differential expression algorithms for DRS data.

Response signatures of Arabidopsis thaliana to single stress and combination of stresses in multiple natural variants

Barah, P.[1,*]; Rasmussen, S.[2]; Suarez-Rodriguez, M.C.[3]; Mundy, J.[3]; Bjørn, N.H.[2]; and Bones, A.M.[1]

[1]Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway -7491
[2]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark -2800
[3]Department of Biology, University of Copenhagen, Copenhagen, Denmark- 2200

*Presenting author: **Pankaj Barah** (pankaj.barah@bio.ntnu.no)

Environmental stress is a key factor to determine the genome regulation, evolutionary history and geographical distribution of living organisms. Plants are sessile organisms and hence unable to escape from unfavourable environments. Complex defense systems have been evolved in plants against different types of stresses. The complex level of network crosstalk makes it challenging to correlate various types of responses to a particular stress type. Some genes involved in stress-response pathways cross-talk with critical physiological processes that are highly conserved among natural variants (ecotypes) and some genes exhibits differential expression behaviour. Such plasticity of the plant transcriptome and its adaptive robustness to respond to complex environmental conditions could be explored further associating SNP information available from 1001 genome project to facilitate a new dimension in stress-resistance engineering.

DNA microarrays is a powerful technique to identify genes exhibiting transcriptional regulation as an effect of changing environmental conditions. As there are many standards on how to grow plants and to conduct transcription experiments, it is difficult to extract and compare information from data sets produced by individual laboratories. To overcome this problem of incompatibility of independent microarray experiments, 23 different genotypes (10 ecotypes and 13 mutants) were subjected to a set of 5 individual stress treatments and 8 combinations of stress treatments under same experimental conditions. This was a part of ERA-NET Plant Genomics, MultiStress project (http://www.erapg.org/). Using a subset of this large dataset we will present some results on-

i) response diversity of seven *Arabidopsis* thaliana ecotypes to a particular stress type (Cold); and

ii) the transcriptional response of two *Arabidopsis thaliana* ecotypes (Landsberg erecta and Colombia) to a combination of two different stresses (cold and high-light).

MELONOMICS: from sequence to natural variation

Sanseverino, W.[1,*] and the MELONOMICS Consortium

[1]Centre de Recerca en Agrigenòmica (CRAG) CSIC-IRTA-UAB-UB Bellaterra 08193 BARCELONA

*Presenting author: **Walter Sanseverino** (walter.sanseverino@cragenomica.es)

The draft sequence of the melon genome (*Cucumis melo* L.) was obtained using a whole genome shotgun approach based on 454 pyro-sequencing complemented by a BAC-end scaffolding. A total of 375 Mb (83.3% of the estimated genome size) have been assembled in 1,599 scaffolds and 29,865 contigs. 87.5% of the genome assembly has been anchored to the melon genetic map. We predicted 27,427 protein-coding genes, which we analyzed by reconstructing 22,218 phylogenetic trees, allowing mapping of the orthology and paralogy relationships of sequenced plant genomes. Two different orthology approaches were used to produce a clear picture of synteny relationships between melon and cucumber, highlighting complex rearrangements between two near-phylogenic species. We observed the absence of recent whole genome duplications in the melon lineage since the ancient eudicot triplication and our data suggest that transposon amplification may in part explain the increased size of the melon genome when compared to the close relative cucumber. Additionally, seven melon accessions have been resequenced to a 20× coverage and mapped to the reference genome, allowing the massive identification of SNPs and INDELs and opening the way for further analysis of structural variation using a reference-guided assembly approach.

SVfinder: an *ab initio* approach for the genome-wide characterization of novel transposable element insertions and other large structural variations

Madoui, MA[1]; Aury, JM[1]; Labadie, K.[1]; Etcheverry, M.[2]; Le Pen, J.[2]; Artiguenave, F.[1]; Wincker, P.[1]; Colot, V.[2]; Gilly, A.[1]

[1] Genoscope, CEA – Universite d'Evry, Evry, France
[2] Institut de Biologie de l'Ecole Normale Superieure, CNRS UMR8197 - INSERM U1024, Paris

*Presenting author: **Arthur Gilly** (agilly@genoscope.cns.fr)

Whole genome sequencing (WGS) has revealed that transposable elements (TEs) are major players in genome evolution and that they make up, together with their relics, a large fraction of many eukaryotic genomes. Documenting by WGS the contribution of TEs to genome variation is however complicated by the repeated nature of these sequences. We developed a method implemented into SVfinder to characterize structural variants (SVs) including duplications, deletions and inversions, TE donors and targets involved in newly transposed TEs in resequenced genomes using paired reads.

SVfinder selects discordant reads pairs directly from the mapping output. Discordant mate-pair reads are clustered using single linkage depending on estimated parameters such as depth coverage and fragment size variation. SV calling is performed by computing mapping signatures.

Efficiency of SVfinder was tested on *Arabidopsis thaliana*. Five kb mate-pair libraries were built from 11 near-isogenic A. *thaliana* lines that likely differ by several new TE insertions (Johannes et al, PLoS Genet, 2010) and each library was sequenced using Illumina technology. SVfinder identified indeed many novel TE insertions in each line, as well as other structural variants.

Key words: structural variations, algorithm, transposable elements, resequencing.

A first attempt towards an epigenetic atlas of the moss Physcomitrella Patens using NGS data

Symeonidi, A.[1,*]; Widiez, T.[2,3]; Luo, C.[3]; Lam, E.[3]; Lawton, M.[3]; Rensing, S.A.[1]

[1]University of Freiburg
[2]University of Geneva
[3]Rutgers University

Presenting author: **Aikaterini Symeonidi** (aikaterini.symeonidi@biologie.uni-freiburg.de)

Chromatin is the natural template of the DNA molecules in the cell. It is formed by histone proteins, on which DNA is wrapped around. The "histone code hypothesis" suggests that the transcription of DNA is partly regulated by chemical modifications to the histone proteins and primarily on the N-terminus of the proteins, suggesting that regulatory information can be derived by histone combinations and post-transcriptional modifications. By using high-throughput sequencing methods on a model organism (*P. patens)* with a known genome, we make a first attempt to investigate the chromatin organization and address the "histone code hypothesis" question in a bryophyte species. Our overall aim in this project is to shed light to the interplay between gene activity and chromatin status in response to dehydration stress as well as developmental stages of the moss *P. patens*.

By using ChIP-seq techniques, we acquired data for five histone modifications (H3, H3K4me3, H3K9Ac, H3K9me2, H3K27Ac and H3K27me3). In total we have 24 libraries, each one corresponding to one histone modification for 3 tissues (protonema, gametophore and gametophore under stress) as well as Input DNA and MOCK DNA for these three tissues, that are used as controls. The 50bp long colorspace data were mapped onto the *P. patens* genome, using BWA, and processed by using established bioinformatics tools (MACS, Bedtools) and pipelines, as well as additional ad hoc programs and scripts when needed. Having acquired the positions of the histone marks within the genome we are in the process of creating a genome-wide map of histone modifications under different conditions and developmental stages. In the next steps, we will conduct a comparison between microarray data with regard to differentially expressed genes under the above-mentioned conditions in order to investigate which marks, under which combinations and how they affect gene expression.

Accounting for variability within ChIP-seq datasets using the R package NarrowPeaks

Madrigal, P.[1,*]; Krajewski, P.[1]

[1]Institute of Plant Genetics, Polish Academy of Sciences, Poznań, Poland

*Presenting author: **Pedro Madrigal** (pmad@igr.poznan.pl)

Background
Next-generation sequencing is enabling the scientific community to make a step further in the understanding of molecular mechanisms controlling transcriptional regulation. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is being applied to globally map transcription factor binding sites for a protein of interest. State-of-the-art computational algorithms for the analysis of this type of ChIP-seq data focus on the detection of the so-called peaks, i.e., enriched regions originated after read sequences are aligned to the reference genome. However, the results produced by the peak callers can substantially differ in number, extension, significance and shape of these peaks. Even with the gain in resolution accomplished by ChIP-seq in comparison to ChIP coupled with microarray technology (ChIP-on-chip), it has been shown recently that there is room for improvement in the algorithms to narrow down the predicted loci and discriminate between true transcription factor binding sites close to each other.

Results
We have developed an R package able to analyze the variability present in a set of transcription factor binding sites returned by other tools, using a trimmed version of functional principal components analysis, to split, narrow down, and filter out a list of candidate peaks. The results show that the variation across a genome-wide read coverage profile can be very informative regarding the actual occurrence of functional protein-DNA interactions. Enabling the detection of narrower peaks at a given locus uncovers a higher concentration of motif presence within the sites.

Availability
http://www.bioconductor.org
http://www.sysflo.eu

An alternative approach to assembly validation highlights deficiencies of the N50 statistic

Bayer, M.M.[1,*]; Marshall, D.F.[1]

[1]Information and Computational Sciences Group, James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland, UK

*Presenting author: **Micha Bayer** (micha.bayer@hutton.ac.uk)

With the advent of Next Generation Sequencing (NGS) a multitude of de novo assembly tools has emerged that deal with the exceptionally difficult task of reconstructing a genome from sometimes very short sequence fragments. The quality assessment of these de novo sequence assemblies is a critical part of modern genomics research, but there is considerable debate over the methodologies used for this.

One widely used metric is the N50, or its derivatives. The N50 is the contig length *L* at which 50% of the bases in the assembly are contained in contigs longer than *L*. There is a widely held belief that larger N50 values, and a lower number of contigs, are indicative of a "better" genomic assembly. However, these statistics provide no indication of the preservation of the gene space in an assembly, i.e. the number of genes that been assembled correctly and exist intact on a single assembled contig. From a biological point of view the preservation of the gene space is critical though, and more relevant than contig size or number.

We have developed a new approach to validating assemblies that puts emphasis on the biologically meaningful assessment of assembly quality. The approach is based on an all-by-all BLAST of a set of full length cDNA (FLcDNA) sequences against all contigs in an assembly. A simple algorithm is then used to examine the BLAST output for integrity of the hits for a given FLcDNA. This tests for whether the entire FLcDNA is covered by a single contig, and counts the number of different contigs hit where this is not the case.

Provisional results seem to indicate that both the N50 and contig number correlate poorly with the number of intact genes, and that gene preservation is in fact more closely related to sequencing depth. Thus, assemblies supported by favourable N50 and contig number statistics may in fact be poor representations of the original genome.

A weighted frequency approach for error detection in NGS metagenomic data

Krachunov, M.K.[1]; Vassilev D.I.[2,*]

[1]Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Bulgaria
[2]Bioinformatics group, AgroBioInstitute, Sofia, Bulgaria

*Presenting author: **Dimitar Vassilev** (jim6329@gmail.com)

Background
Because of the way metagenomic samples are collected and sequenced, it is not possible to re-sample the same data again as a way to account for any sequencing errors. This not only increases the difficulty in distinguishing biological variations from equipment errors, but the unavoidable noise in the data inevitably leads to significant changes in the results of any metagenomic experiment. For improving of the quality of such studies, the development of an optimal error detection and correction approach is required.

There are many published methods for NGS error detection, but most of them focus on the task of improving de novo genome sequencing, so they are not designed to work with metagenomic data sets or take into account the nature of metagenomic data and its inherent characteristics.

Results
The proposed approach uses base frequency counting to estimate the errors, but it introduces a weight that gives focus to equal and similar reads, thus accounting for multiple reads representing the same sequence as well as for related sequences, while penalizing any completely unrelated ones. As the same genetic sequence is inherited by multiple organisms in the sample with varying degrees of conservation, related organisms might provide additional clue for the correctness of the sequence, but with a low enough degree of significance so that real biological variation doesn't get filtered out.

Sequences of different organisms aren't generally counted as additional coverage for the genome of a given organism, but as they increase the information known for it, their utilization could lead to an improvement of the quality in areas with low coverage.

Conclusions
The biggest challenge in developing such approach is the task of evaluating it. Due to the difficulty in obtaining large quantities of confirmed or corrected metagenomic sequences it's not clear how to estimate the change in false positives and false negatives with certainty. One currently utilized approach is to crudely estimate the errors introduced by the equipment and to try to simulate similar errors in sequences that are presumed to be correct.

Patterns of nucleotide asymmetries in plant and animal genomes

Mascher, M.[1,*]; Schubert, I.[1]; Scholz, U.[1]; Friedel, S.[1]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany

*Presenting author: **Martin Mascher** (mascher@ipk-gatersleben.de)

Background

Symmetry in biology provides many intriguing puzzles to the scientist's mind. Chargaff's second parity rule states a symmetric distribution of oligonucleotides within a single strand of double-stranded DNA. While this rule has been verified in a wide range of microbial genomes, it still awaits explanation.

Results

In our study, we inquired into patterns of mono- and trinucleotide intra-strand parity in complex plant genomic sequences that became available during the last few years, and compared these to equally complex animal genomes. The degree and patterns of deviation from Chargaff's second rule were different between plant and animal species. We observed a universal inter-chromosomal homogeneity of mononucleotide skews in coding sequences of plant chromosomes, while the base composition of animal coding sequences differed between chromosomes even within a single species. We also found differences in the base composition of dicot introns in comparison to those of monocots. These genome-wide patterns were limited to genic regions and were not encountered in inter-genic sequences.

Conclusions

Our findings provide new insights as to the intra-strand parity of nucleotide distribution in the large and complex genomes of several plant and animal species and shed new light on various hypotheses about functional correlations of intra-strand parity which have hitherto been put forward. Furthermore, we propose more recent polyploidization and subsequent homogenization of homoeologues as a possible reason for more homogeneous skew patterns in plants.

A fully automated pipeline for the analysis of Liquid Chromatography-Mass Spectrometry (LC-MS) based metabolomics experiments

Franceschi, P.[1,*]**; Scholz, M.**[1]**; Shahaf, N.**[1]**; Wehrens, R.**[1]

[1]Biostatistics and Data Management, Fondazione Edmund Mach

*Presenting author: **Pietro Franceschi** (pietro.franceschi@fmach.it)

Background

The recent improvement in analytical technologies has been the ground for the advent of high throughput metabolomics. This member of the "omics" family has the objective of fully characterizing biological systems at a comprehensive metabolic level. As in many other "omics" cases, however, metabolomics datasets show a high level of complexity and the development and the optimization of dedicated data preprocessing and analysis tools is of paramount importance to guide biological interpretation and biomarker identification.

Methods

In this communication we will present the fully automated data analysis pipline for LC-MS based metabolomics, which have been recently set-up at the FEM Research and Innovation Centre. The pipeline has been developed in R and aims at a seamless integration of data preprocessing, quality assessment, feature annotation and data analysis within a unified framework. The pipeline has been integrated in a web based application which can be directly used by the scientists lacking a specific bioinformatic background.

Data preprocessing and feature extraction is performed by the widely diffuse xcms package, running with a set of parameters tailored on the technological platform installed at FEM. After this step, the assessment of the data quality relies on set of visualization tools based on Principal Component Analysis. As a third step, experimental features are "annotated" with the objective of assigning them a "chemical" and "metabolic" identity. Since this last step represents undoubtedly the critical stage in metabolomics, a strong effort has been put into the annotation modulus, both by implementing annotation against an in-house developed database and also by developing a new tool to adaptively calculate the mass tolerance used for database search.

Results

Even if the pipeline is still under active development, preliminary results clearly indicates that:

1) R can be used as an effective environment to develop data analysis tools suitable for the use by a wide scientific community;

2) the solutions implemented to perform quality control allow the fast identification of critical/bad samples and also the early identification of drifts in the analytical pipeline;

3) the implementation of the adaptive mass tolerance window for database search guarantees an improvement in the quality of the annotation both in terms of a reduced number of false positive and of a better identification of low concentration compounds.

VERONA
USEFUL INFORMATION

PRACTICAL INFORMATION (A-Z)

BANKS
In the city centre there are several banks open from 8.30 am to 12.30 am and from 2.30 pm to 4.00 pm approximately. There are many ATMs 24 hours a day - 7 days a week.

CREDIT CARDS
In Italy all the main credit cards are accepted.

CRIME AND PERSONAL SAFETY
Verona centre is safe, nevertheless we recommend the use of common sense, keeping an eye on wallets when in crowds and possibly avoiding carring too much cash or wearing flashy jewellery.

CURRENCY
The Italian currency is Euro (EUR o €)
The approximate exchange rate (December 2011) is: 1 EUR = 1.3028 US $; 0.8395 £

ELECTRICITY
Italian voltage is 230 and frequency is 50 Hz.
The electric plugs are mainly type L (three pole "Italian" plug). This standard includes two models rated at 10 A and 16 A that differ in contact diameter and spacing. Both are symmetrical, allowing the plug to be inserted at either direction.

LANGUAGE
The official language is Italian. Verona is a popular tourist destination, so most of the restaurants and bars staff speak English and German.

LOCATION AND WEATHER
Verona, located in the north east part of the Italian peninsula, is 59 meters above sea level. The city experiences hot summers and cold, humid winters, even though tempered by Lake Garda's influence.

OPENING HOURS
Offices
Monday - Friday 9.00 am/1.00 pm - 2.00 pm/6.30 pm
Shops
Monday 3.00 pm / 7.30 pm - Tuesday/Saturday 9.30 am/7.30 pm
(some shops open all day from 9.30 am to 7.30 pm)
Supermarkets
Monday - Friday 8.30 am/12.30 pm - 3.30 pm/7.30 pm
In the city centre, supermarkets usually stay open from 8.30 am to 7.30 pm or later.

SMOKING
On 10th January 2005 smoking bans were introduced in all the public place, such as bars, restaurants, hotels, offices, shops, etc. with the exception of the places with reserved smoking areas, provided with a suitable ventilation system.

TAXES
In Italy all the services and goods are usually subject to VAT tax that is 21% of the value of the goods. In restaurants VAT is 10% and it's usually already included in the bill.

TELEPHONES
Italy country code is **0039**
For international calls, dial **00 + national code + area code + personal number**

TIME

The Italian local time is UTC (GMT) + 1, that means one hour ahead of Greenwich mean time. Italy adopts the daylight saving time (DST) from the last Sunday of March till the last Sunday of October.

TIPPING

In all restaurants, bars, taxis etc. the service is included in the price. Tips are happily accepted but it is not necessary to leave anything more than the amount assessed.

EMERGENCIES

In case of sanitary **emergencies or dangerous situations** the number to call is **118**
Carabinieri (Italian military police) **112**
Police 113
Fire brigade 115

CHEMIST'S

In the city centre there are several chemist shops, easy to recognize by a bright green cross.
They open Monday - Friday from 9.00 am to 12.00 am and from 4.00 pm to 7.00 pm.
Duty pharmacies open also on weekends and holidays. Find the list of the chemist shops with the timetable on www.farmacieverona.it or outside every chemist shop.

HOSPITALS

Ospedale Civile Maggiore
Piazzale A. Stefani, 1
37126 - Verona

Policlinico G. Rossi
Piazzale L.A. Scuro, 10
37134 - Verona

Switchboard 045 812 11 11

GETTING AROUND

Verona city centre is extremely rich in historic and architectural beauty; so it is certainly worth a visit by foot. This is the only way to enjoy completely the view of magnificent palaces, squares and monuments that bring us to an imaginary trip back to Verona's glorious history. Moreover, enjoy picturesque glimpses that stimulate your attention and enliven your fantasy!

CARS

Be careful about driving into Verona city centre because of ZTL areas, which are limited traffic flow areas where the access of cars is limited to pre-established times.

FREE ACCESS TO THE ZTL TIMETABLE

Monday - Friday:
- from 10.00 am to 1.30 pm
- from 4.00 pm to 6.00 pm
- from 8.00 pm to 10.00 pm

Saturday, Sunday and holidays: - from 10.00 am to 1.30 pm
If staying at a hotel inside a ZTL area, make sure the hotel or its garage calls the licence numer into the police and optains a provisional transit permit. Keep the hotel bill in case you need to challenge a fine later.

PARKING
Free parking:
- Porta Palio Parking - stradone Porta Palio
- Piazzale Guardini Parking - piazzale Guardini
- Stadium Parking (three park and ride) - piazzale Olimpia

Paying car parks:
- Arena Parking - via Bentegodi, 1 - tel. 0039 045 8009333
- Cittadella Parking - piazza Cittadella, 4 - tel. 0039 045 595593
- Italia Parking - corso Porta Nuova, 91 - tel. 0039 045 8006312
- Piazza Isolo Parking - via Ponte Pignolo, 6 - tel. 0039 045 8007921
- Via Città di Nimes Parking - via Città di Nimes - tel. 0039 045 2320025*
- Arsenale Parking - via Arsenale 8 - tel. 0039 045 8303460
- Passalacqua Parking - via dell'Università - tel. 0039 045 2320025*
  It is the closest to Conference venue, just in front of the Polo Zanotto. For this open air parking place you can pay inside when leaving it. "Passalacqua" parking is open 24 hours a day-7 days a week. It is located in Viale dell'Università about 100 metres far from Polo Zanotto.

Parking is also allowed along certain roads in the city, in areas with white and blue painted control lines. In the second case drivers must use a special prepaid coupon called "Verona Park", on sale at tobacconists and paper shops. For further information www.comune.verona.it

VERONA CARD
VeronaCard is an all-inclusive ticket that allows to gain free entry to museums, churches and monuments in the city and travel for free on ATV bus services.
There are two cards available:
- VeronaCard valid for two days at the price of € 15,00
- VeronaCard valid for five days at the price of € 20,00

You can buy VeronaCard at museums, monuments, churches, tobacconists and tourist information point in the city centre and at Garda lake surroundings and at all the sales points which participate in the initiative.
For further information click http://www.comune.verona.it/turismo/veronacard.htm

PUBLIC TRANSPORT
Azienda Trasporti Veronesi (ATV) runs urban public transport. Tickets, valid within 60 minutes from the validation, can be bought at tobacconists for € 1,10 or € 1,50 on board. For € 3,50 it is possible to buy a daily ticket, valid 24 hours long on all urban routes.
Here is a synthesis of the main routes in the city:
- ***from the station to piazza Bra (Arena)***
  Monday - Saturday: buses 11, 12, 13, 72; on Sunday: 90, 92, 93, 96, 97
- ***from the station to Castelvecchio***
  Monday - Saturday: 21, 22, 23, 24, 41; on Sunday: 91, 93, 94, 95
- ***from Piazza Bra to Polo Zanotto (Conference Venue)*** **Via XX Settembre bus stop**
  Monday - Saturday: 11, 12, 13, 510; on Sunday: 90, 92, 98, 510*
  *Only in the afternoon
- ***from Station to Polo Zanotto (Conference Venue)*** **Via XX Settembre bus stop**
  Monday - Saturday: 11, 12, 13, 51, 510; on Sunday: 90, 92, 98 and 510* **platform A**
  * Only in the afternoon

TAXI
Radiotaxi switchboard (24 hours a day - 7 days a week) 0039 045 532 666
Taxi stand Piazza Bra +39 045 803 0565
Taxi stand Porta Nuova Station +39 045 800 4528

THE CITY OF VERONA

Verona is one of the most important cities in Veneto. It is well known for its artistic and architectural beauties. Thanks to its history the city centre was listed by UNESCO as a World Heritage Site.

ONE OF THE POSSIBLE ITINERARY
Starting from Verona railway station, along Corso Porta Nuova: you will easily reach the heart of the city, Piazza Bra, with the monumental Arena. The Romanic Anfiteatro is one of the most ancient and best-preserved of those existing, and it's now the charming venue of a famous Opera season. Give at least a glance at the great palaces Gran Guardia and Barbieri and take for the Liston, a characteristic wide pavement on that goes through piazza Bra and brings to via Mazzini, a street with plenty of shops, bars and restaurants. On arrival at the end of via Mazzini you will be surprised to find a very peculiar dwelling: the very famous Casa di Giulietta (Juliet's house). At this point you will certainly have a feeling for the city fascinations: the inner town streets hold unexpected sightsand ancients buildings, churches, and dramatic views of bridges and monuments.
A few more steps and you will find yourself in Piazza delle Erbe, the square where the city market has been taking place for centuries. The square is surrounded by some extraordinarily interesting buildings: the Domus Mercatorum, the Gardello Tower, Maffei Palace and Case Mazzanti (a huge completely painted "a fresco" house). Pass over the Arco della Costa, and see the beautiful Piazza dei Signori, that used to be, together with the adjacent Piazza delle Erbe, the venue of government palaces in Middle ages and it's now a very appreciated meeting place, where social and cultural events are set almost every day. Very close to Piazza dei Signori are the Arche Scaligere, monumental graves of Signori della Scala, Lords of Verona straddling the XIII and XIV centuries, situated in front of the Church of Santa Maria Antica.

VERONA CITY SIGHTSEEING
With Verona City sightseeing you will have the opportunity to join the tour at any stop, listen to the commentary in one of the seven languages available or reach any other bus-stop on the route and see all the best sights and attractions that the city offers.
There are two routes available, Line A and Line B: choose the one that more stimulate your curiosity!
Ticket is valid 24 hours on both itineraries.
Hop off whenever you like and hop on every 60 minutes.
Wheelchairs are welcome.

FARES
Adults: € 18.00
Children 5-15 years old: € 9.00
Family (2 adults + 2 children): € 50.00
Family (2 adults + up to 3 children): € 55.00
For further information www.verona.city-sightseeing.it

CHURCHES AND MUSEUMS
In Verona there are several churches of great historic and artistic value. The most famous is the Basilica dedicated to San Zeno, held as a Romanesque masterpiece. Rather inspired by Gothic is Sant'Anastasia Church, which is home to some famous Pisanello's and Altichiero's "a frescos". Remarkable too, the Dome (Santa Maria Matricolare Cathedral) and San Fermo Church, rare because of the four naves division, in its lower part.

Following is a short overview on the main museums:
- ***Castelvecchio Museum***

It is in Castelvecchio fortress, an imposing building of the Middle Ages. The exhibition is hosted in about thirty rooms and it's divided into sculpture, Italian and foreigner painting, ancient weapons, ceramic ware, jewels, miniature paintings.

The old town bells are also on view.

- ***Jiuliet's grave and Museo degli Affreschi***

The museum is hosted inside the monastery of San Francesco al Corso. Visitors can admire cycles of frescos from Veronese medieval buildings in the sixteenth century, and nineteenth century sculptures, while the church of San Francesco hosts works on canvas of great dimensions ranging from the sixteenth to the eighteenth century. Inside, a crypt, under the church of San Francesco, lies an empty, simple sarcophagus made of red Verona marble. It is believed to be the grave where finally rested the legendary Veronese Lovers.

- ***Museo Lapidario Maffeiano***

This is the oldest public museum in Europe. It hosts an important collection of memorial and funeral epigraphs.

- ***Teatro Romano and Museo Archeologico***

In the numerous exhibition spaces a large range of archaeological finds is displayed, like sculpted and decorative pieces coming from the Romanic Theatre, mosaics, votive and funeral stones, altars, epigraphs, sculptures, glass, pottery etc. Entrance ticket to the Museum includes also the visit to the Theatre.

SHOPPING

Via Mazzini and via Porta Borsari are appreciated and well known for the great deal of beautiful shops they display.

Footwear, clothes, underwear, leather goods stores, etc. all the best that Italian high fashion labels offer to fashion victims!

EATING AND DRINKING

PUBS & CLUBS SUGGESTED

PUBS & CLUBS	ADDRESS	PHONE	CLOSING DAY (always verify)	PUB OR CLUB	AVERAGE PRICE
Alla Cappa	P. Bra Molinari 1	0458004516	-	Cocktail	€
Filippini	P. Erbe 26	0458004549	Wednesday	Cocktail	€
Square	Via Sottoriva 15	045597120	Monday	Cocktail & trends	€
Le Cantine de l'Arena	Piazzetta Scalette Rubiani, 1	045 8032849	-	Music Brasserie	€
K59	Via Carlo Montanari, 10	045 8015650	-	Cocktail & trends	€

RESTAURANT SUGGESTED

RESTAURANT	ADDRESS	PHONE	CLOSING DAY (always verify)	CUISINE	AVERAGE PRICE
Bella Napoli	Via Guglielmo Marconi, 16	045 591143	-	Pizzeria	€
Pizzeria Leone "da Ciro"1924	Via Zambelli Giovanni, 20	045 806 5161	Monday	Pizzeria	€
Trattoria Alla Colonna	L.go Pescheria Vecchia 4	045596718	Sunday	Typical	€€
Hostaria La Vecchia Fontanina	P.tta Chiavica 5	045591159	Sunday	Typical	€€
Osteria Carroarmato da Annalisa	Vicolo Gatto 2	0458030175	Monday	Osteria and food	€€
Osteria Sottoriva (Franco)	Via Sottoriva 9/a	045 8014323	Wednesday	Osteria and food	€€
Osteria dell'Orso	Via Sottoriva 3/c	045597214	Sunday/Monday at lunch	Osteria and food	€€
Osteria Dal Duca	Via Arche Scaligere 2/b	045594474	Sunday/ Tuesday at lunch	Typical	€€
Trattoria Giulietta e Romeo	Corso Santa Anastasia 27	0458009177	Sunday/ Monday at lunch	Typical	€€
Trattoria Da Ugo	Vicolo Dietro S. Andrea 1/b	045594400	Sunday /Monday at lunch	Typical	€€€
Trattoria I Masenini	Via Roma 34	0458065169	Sunday	Typical	€€€/€
Trattoria Pescheria I Masenini	P.tta Pescheria 9	0459298015	Sunday/Monday at lunch	Fish	€€€€
Trattoria Al Calmiere	P.zza S. Zeno 10	0458030765	Sunday evening/ Monday	Typical	€€€/€
Ristorante Antico Caffè Dante	Piazza dei Signori	0458000083	Sunday evening/Monday	Refine	€€€€
Ristorante Maffei	P. Erbe 38	0458010015	-	International	€€€€
Ristorante Maria Callas	Vic. S. Pietro Incarnario	045594034	Sunday	Refined	€€€/€
Trattoria di Giovanni Rana	Piazza Brà 16	0458002462	Sunday evening/ Monday	Refined	€€€€/€
Ristorante Ai Teatri	Via S. Maria Rocca Maggiore	0458012181	Sunday/ Monday at lunch	Refined	€€€/€
Locanda Castelvecchio	Corso Castelvecchio 21/A	0458030097	Tuesday/Wedn. at lunch	Typical	€€€€/€
Ristorante Re Teodorico	P.le Castel San Pietro	0458349990	Wednesday/ Sunday evening	Refined	€€€€
Ristorante Dodici Apostoli	V.Lo Corticella S. Marco 3	045596999	Sunday evening/Monday	Refined	€€€€/€
La Bottega del Vino	V. Scudo Di Francia 3	0458004535	Tuesday	Typical (Wines)	€€€€/€
Trattoria Al Pompiere	V.lo Regina d'Ungheria 5	045 8030537	Sunday / Monday at lunch	Typical	€€€€
Ristorante La Fontanina	P.tta Portichetti Fontanelle 3	045913305	Sunday/ Monday at lunch	Refined and Romantic	€€€€€
Ristorante Il Desco	Vic. S. Sebastiano 3/5	045595358	Sunday/ Monday	Refined	€€€€€
Osteria Oste Scuro	V.lo S. Silvestro, 10	045592650	Sunday / Monday at lunch	Fish	€€€€€

€ = 10/25 EURO	€€ = 25/35 EURO	€€€ = 35/40 EURO	€€€€= 40/50 EURO	€€€€€ = more than 50 EURO

TOURISM

GARDA LAKE

If the weather is fine, and you need to take a break from business, there is nothing better than a journey to the lake. The villages around Garda lake are popular tourist destinations, visited every year by hundreds of thousands of people.

The surrounds of the lake are ideal for several activities: the stretch of water is perfect for windsurf, canoeing, scuba diving or having fun on a recreational boat. Alternatively, you can just sunbathe or have an invigorating swim.

The next up-country is also good for cycling, golf, horse-riding and much more. Moreover in proximity of the lake there are several natural reserves, amusement and water parks.

www.lagodigarda.it

LESSINIA

Lessinia landscape is made very special by its natural beauties and historical statements. It is a wide highland covered with woods and pastures, located for the most part in the provinces of Verona and Vicenza. The countryside has exceptional geological features and it is also rich in a large variety of rare plants. There you can also find picturesque towns with stone-roofed houses, made by the former inhabitants of this area.

In 1990 it was established the Regional Nature Reserve of Lessinia, created with the aim to protect its historical, archaeological and natural richness, and to preserve all the aspects of a very much appreciated destination for excursions on foot or by cycle.

www.lessiniapark.it

VALPOLICELLA

Valpolicella, famous for the production of its excellent wine (Valpolicella, Recioto della Valpolicella and Amarone above all), is geographically a body of small valleys, gently sloping towards the hills behind Verona. Blessed with a mild climate, the landscape is made up of low rolling hills, filled with vineyards, olive and cherry tree groves, and passed through by widespread accesses.

Besides, a high interest is raised by the many numerous villas, ancient churches and small stone built villages.

www.valpolicella.it / www.valpolicellaweb.it

SOAVE

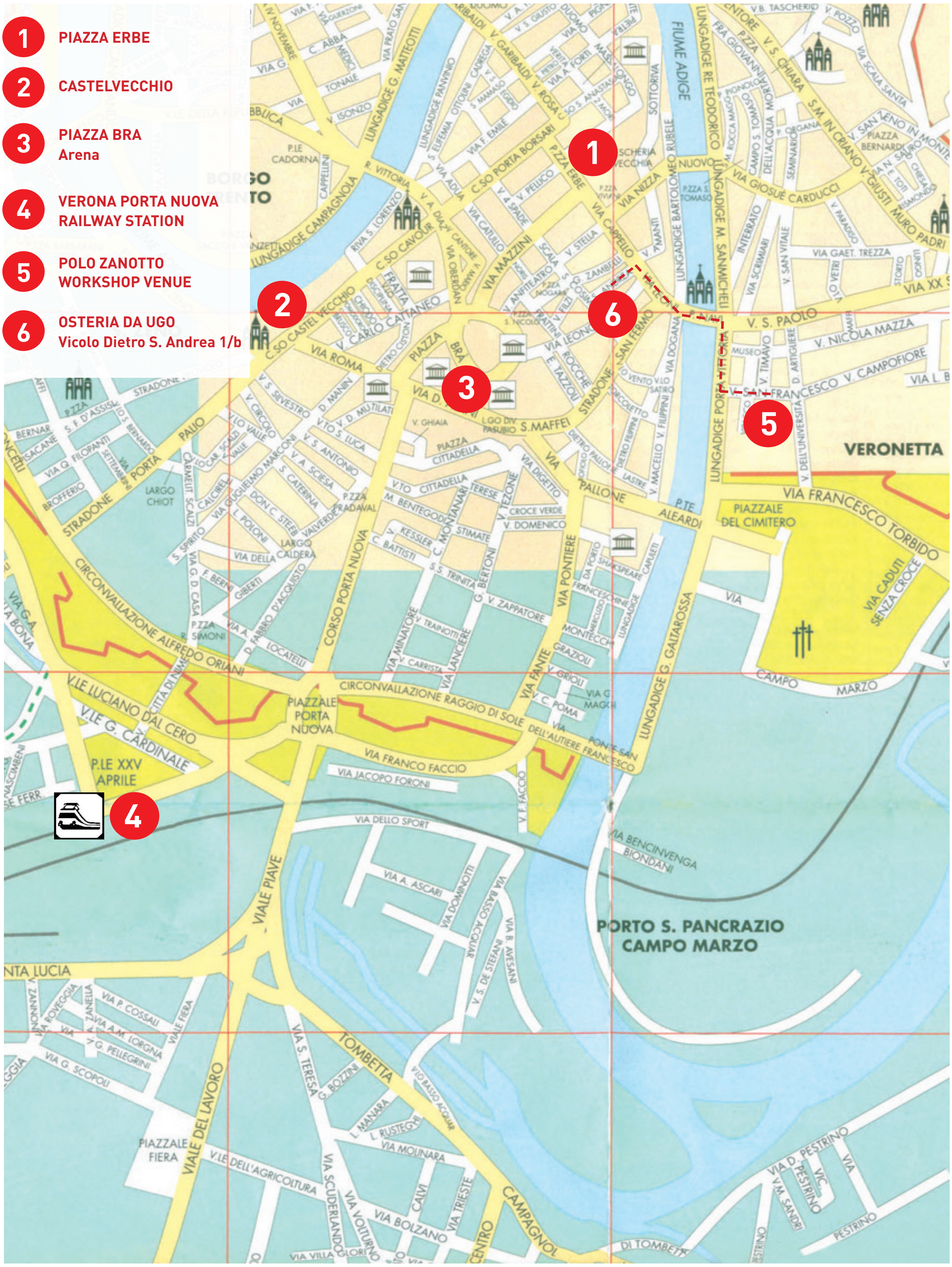
The Soave area is well-known for its wines: Recioto di Soave (a white whine) in particular has been the first wine to obtain DOCG certification (a special mark guaranteeing its high quality and its link with the origin area). It is a classic dessert wine, with digestive properties, ideal for dining accompaniment such as cream desserts, pastries (as “Pandoro di Verona” cake), blue cheeses and pâté.

The territory has also an historic importance. The town still maintains a middle-age structure, with an intact boundary wall climbing up the hill above, where the Scaligero Castle stretches skywards.

In September, on the occasion of the grape harvest, many events like grape fests or historical re-enactments, are scheduled.

www.comunesoave.it

MAP OF VERONA


1 PIAZZA ERBE
2 CASTELVECCHIO
3 PIAZZA BRA Arena
4 VERONA PORTA NUOVA RAILWAY STATION
5 POLO ZANOTTO WORKSHOP VENUE
6 OSTERIA DA UGO Vicolo Dietro S. Andrea 1/b
VERONETTA
FIUME ADIGE
PORTO S. PANCRAZIO CAMPO MARZO
PIAZZALE PORTA NUOVA
CORSO PORTA NUOVA
VIA ROMA
VIA MAZZINI
CIRCONVALLAZIONE RAGGIO DI SOLE
CIRCONVALLAZIONE ALFREDO ORIANI
VIA FRANCESCO TORBIDO
PIAZZALE DEL CIMITERO
P.LE XXV APRILE
VIALE PIAVE
VIALE DEL LAVORO
PIAZZALE FIERA
LUNGADIGE PORTA VITTORIA
LUNGADIGE G. GALTAROSSA
V. S. PAOLO
VIA XX SETTEMBRE