CONICYT
Ministerio de Educación
Gobierno de Chile

Comisión Nacional de Investigación Científica y Tecnológica – CONICYT

FONDAP CENTERS OF EXCELLENCE IN RESEARCH PROGRAM

ANNUAL PROGRESS REPORT

Center for Genome Regulation (CGR)

2013

cgr Center for Genome Regulation

I.

PRESENTATION

PERIOD REPORTED: 1st Year ☐ 2nd Year ☐ 3rd Year ☐ 4th Year ☐ 5th Year ☐

PERIOD COVERED: From: January 1, 2013 To: December 31, 2013

NAME OF THE CENTER **FONDAP Center for Genome Regulation (CGR)**		**CODE** 15 09 00 07
DIRECTOR OF THE CENTER Miguel L Allende	**E-MAIL** **allende@uchile.cl**	**SIGNATURE**
DEPUTY DIRECTOR Martín Montecino	**E-MAIL** **mmontecino@unab.cl**	**SIGNATURE**
SPONSORING INSTITUTION Universidad de Chile		
ASSOCIATED INSTITUTION(S) (if applicable) Universidad Andrés Bello; Pontificia Universidad Católica de Chile		
CENTER WEBSITE ADDRESS www.genomacrg.cl		

DATE: 31/01/14

RESEARCH LINES

N°	Research Line	Objective	Principal Researcher	Associated Researcher(s)
1	Extreme Genomes: Plants	Comparative genomics of desert plants inhabiting an altitudinal gradient and of sporadic flowering plants of the extremely dry Atacama desert.	Rodrigo Gutiérrez Ariel Orellana	Andrea Miyasaka
	Extreme Genomes: Animals	Analysis of the genomic structure of the *Orestias* fish species and transcriptomic profile of *Austrolebias* fish and of populations of *Rhinella* frogs.	Miguel Allende Martín Montecino	Alvaro Glavic Christian Hödar
	Extreme Genomes: Microbes	Metagenomics of microbes associated with the soil of high altitude plants; Genomes of *Sulfobacillus* sp. and of biomining bacteria.	Mauricio González Alejandro Maass	Verónica Cambiazo
2	Relevant Genomes: *Homo sapiens*	The genome of the Mapuche, one of Chile's indigenous peoples	Rodrigo Gutiérrez Alejandro Maass Mauricio González Martín Montecino Ariel Orellana Miguel L Allende	Juan Francisco Miquel Giancarlo de Ferrari Silvana Zanlungo
	Relevant Genomes: *Vitis vinifera*	The genome of the table grape (sultanina variety)	Ariel Orellana Alejandro Maass	Andrea Miyasaka
	Relevant Genomes: *Salmo salar*	Structure and annotation of the genome of Atlantic salmon	Alejandro Maass Miguel Allende	Verónica Cambiazo

	Relevant Genomes: *Piscirickettsia salmonis*	The genome of this important fish pathogen	Alejandro Maass	Verónica Cambiazo
	Relevant Genomes: *Prunus persica*	Genomics, transcriptomics and proteomics of the peach	Ariel Orellana	Andrea Miyasaka
3	Gene Expression in Cells: Epigenetic mechanisms	Control of genes by chromatin modification and regulatory RNA molecules	Martín Montecino	Verónica Palma
	Gene Expression in Cells: Development, stem cells and regeneration	Molecular biology of differentiation, cell migration and tissue morphogenesis in development and regeneration	Miguel Allende	Verónica Palma Tomás Egaña Alvaro Glavic Giancarlo de Ferrari Christian Hödar
	Gene Expression in cells: The stress response.	Genomic and proteomic outcomes induced by a biotic or abiotic stressor	Ariel Orellana	Alvaro Glavic
	Gene Expression in Cells: Networks and modeling	Use of *omic* data to construct theoretical interaction networks	Alejandro Maass Mauricio González Martín Montecino	

EXECUTIVE SUMMARY

In 2013, the FONDAP Center for Genome Regulation (CGR) has firmly established itself as the leading center for genome science in Chile, being recognized as such both within and outside the scientific community. This has happened because of the maturation of our core projects and because these have become widely known through our publications and media reports. Thus, we are on the way to fulfilling our primary objective: to establish a legacy and a conceptual association of the CGR to a field of scientific and social importance in Chile. This strategic aim has been achieved by focusing our main effort towards the so called "Center Projects", work that congregates many of the Principal and Associated investigators of the CGR in joint efforts. These are mostly genome sequencing projects that involve interesting and relevant biological problems that our country offers. As we narrated in the previous year's report, the genome projects are divided in two large areas. First, we have undertaken the task of examining genomes of species of diverse phyla that inhabit one of the most extreme environments on earth: the Chilean high-altitude desert, or *altiplano*. During 2012, we embarked on several collection expeditions to this remote area and obtained samples of plants, animals and microorganisms present in the environment. During 2013, we prepared the material and proceeded with next generation sequencing strategies and subsequent bioinformatic analysis, which have produced initial results for most of the projects. Some have advanced to the point of manuscript submission; others, are still at the early stages of analysis. A second important group of projects is related to species of economic importance, such as fruit crops, salmon and fish pathogens. Again, articles describing findings in these organisms have been published or have been submitted. We also have completed the first stage of what we consider the most important challenge for the CGR during the first five year period: the sequencing of the Chilean Human Indigenous Genome. A manuscript has been submitted with our findings and we believe it will be a paper of high scientific and social impact. In early 2014 we will be making a public announcement to the media regarding this work and it will most definitely be one of the highlights of our activities. Finally, the third objective of the project which involves more traditional experimental and theoretical molecular biology has continued with strong productivity and new discoveries.

As was the case for the Human Genome Project, other center projects have also benefited from our trademark *Jamboree* method of collaborative work. Close interaction of PIs, biologists and bioinformaticians has allowed the emergence of advances that would be impossible or much more inefficiently achieved without it. We have systematically approached each genome project in this fashion, allowing us to have a proprietary pipeline for preparation, analysis and interpretation of the data. Besides the submission of the first article on the Chiean Human Genome, we have also sent in a manuscript describing the transcriptome of the native frog *Rhinella spinulosa*, where we compare gene expression among populations that inhabit very different environments. Closely following these will be work on the *Orestias* fish, desert plant genomes and desert metagenomes. Importantly,

these articles are, and will be, co-authored by multiple CGR investigators, as part of our policy to stimulate collaborative, rather than individual work. Published during 2013 were articles describing the peach genome and the table grape genome, both reports picked up by the national press. In total, we published 37 papers this year which translates into a steady increase in our productivity as *bona fide* CGR projects become published. Importantly, we show a very trong increase in the proportion of articles co-authored by CGR investigators (4 of 35 in 2012 to 8 of 37 in 2013) and co-directed theses (from 4 to 7), demonstrating that our strategy for stimulating collaboration within the center has succeeded.

Conferences, courses and training.
The number of training and networking activities generated in 2013 has been very high. We organized 12 courses or workshops that had an impact on over 300 scientists and many more young investigators. We held three activities organized jointly by the CGR and foreign centers. One, was carried out in cooperation with the German DFG and allowed the visit of 25 german scientists from the field of stem cell therapies and regeneration; a second event was carried out together with researchers from the University of Tokyo, who also traveled to Chile as part of a large delegation of Japanese visitors. Thirdly, we held a two week course and symposium jointly with the University of Heidelberg and Santander Universities, with 30 foreign guest speakers. All of these instances have strengthened our international ties and have resulted in funding applications for new collaborative ventures.

Outreach
2013 was a very intense year for outreach activities. We organized many events aimed at high school teacher training in molecular and cell biology and we had an impact on children from many regions of Chile. We have formed a vibrant network of school teachers and, with them, we have established long-term projects to follow up with their students. We have also provided support in terms of travel fellowships for outstanding children, equipment and infrastructure for school labs and even the participation of a group of school students in a professional scientific meeting with poster presentation included. Furthermore, our visibility in the public sphere was augmented with frequent appearances of our researchers in national media.

As future challenges for the CGR, the main one is to generate the first wave of high-impact publications that continue to report on the findings we are achieving with our genome projects; these should peak in 2014 and will be followed up with more specific work on the biological rammifications of the data. A second important strategic aim for this year is to emphasize the formation of trained scientists and engineers in advanced bioinformatics. We have noticed a crippling deficiency in the number of bioinformaticians who can handle the types of projects we carry out, a problem that will increase over time as more and more

institutions require this type of expertise. To reverse this trend, we have set a goal to hire and train a significant number of young investigators in this area to satisfy our internal demand for data handling and also as a way to contribute with advanced human capital in a relevant area to the country. We foresee that the next few years will involve a "return to biology" in the field of genomics, as sequencing capacities and data generation are no longer limiting. Rather, we will now have to make sense of the data and generate new hypotheses to test in the organisms that we have chosen to study.

II. ADMINISTRATIVE ASPECTS

1. **Budget execution**: Describe and justify any budgetary modifications (itemized) of the original proposal.

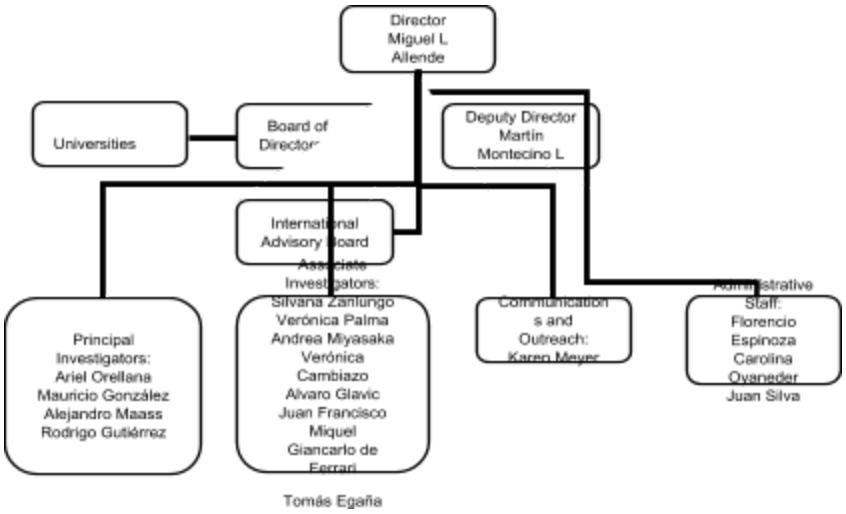
   We have made only minor modifications to this years' budget.

2. **Accomplishment of institutional commitments:** describe any difficulty (ies) encountered regarding this aspect.

   All institutions involved in the project have complied with their obligations in terms of monetary and material support for the investigators. The few exceptions have been satisfactorily resolved.

3. **Organizational Chart:** Present an organizational chart of the Center depicting its main links to companies, associated institutions, and other units within the same institution.

A chart showing the internal organization of the CGR



A chart showing principal institutions interacting with the CGR

4. Personnel

Table indicating position and hourly commitment of all CGR personnel (scientific and administrative) regardless of their funding source involved with the CGR during 2012. A commitment of 44 hours refers to the weekly commitment during the period in which they belonged to the Center.

Name	Position at CGR	Commitment (hours)
Miguel L Allende	Director	44
Martín Montecino	Deputy Director	30
Ariel Orellana	Principal Investigator	26
Alejandro Maass	Principal Investigator	26
Rodrigo Gutiérrez	Principal Investigator	26
Mauricio González	Principal Investigator	26
Silvana Zanlungo	Associate Investigator	15
Verónica Palma	Associate Investigator	15
Andrea Miyasaka	Associate Investigator	15
Verónica Cambiazo	Associate Investigator	15
Alvaro Glavic	Associate Investigator	15
Juan Francisco Miquel	Associate Investigator	15
Tomás Egaña	Associate Investigator	15
Christian Hödar	Associate Investigator	15
Nicolás Loira	Post doc	44
Vicente Acuña	Post doc	44
Phillippe Bordon	Post doc	44
Rodrigo Pulgar	Post doc	44
Dinka Mandakovic	Post doc	44
Leonardo Pavez	Post doc	44
Fernan Federici	Post doc	44
Rosario Villegas S	Post doc	44
María Laura Ceci	Post doc	44
Julian Verdonk	Post doc	44
Macarena Vargas	Post doc	44
Laura Gallardo	Post doc	44
Paula Vizoso	Post doc	44
Giorgia Daniela Ugarte	Post doc	44
Elena Vidal Olate	Post doc	44
Diana Grass	Post doc	44
Andrea Vega	Post doc	44
Adriana Batias	Post doc	44
Henriett Pál-Gábor	Post doc	44
Javier Canales	Post doc	44
Ann Reckhenrich	Post doc	44

Rodrigo Pulgar	Post doc	44
Leonardo Pavez	Post doc	44
Christian Hodar	Post doc	44
Alejandro Zuñiga	Post doc	44
Talia del Pozo	Post doc	44
Rodrigo Assar	Post doc	44
Gino Nardocci	Post doc	44
Felipe Veloso	Post doc	44
Dr. Luis Milla	Post doc	44
Catalina Prieto	Post doc	44
Jose Antonio O`Brien	Post doc	44
Luisa Pereiro	Post doc	44
Henriet Pal`Garbor	Post doc	44
Karina Castillo	PhD student	44
Andrés Aravena	PhD student	44
Alexander Frank	PhD student	44
Sebastián Donoso	PhD student	44
Adrián Moreno	PhD student	44
Marcelo Alarcon Lozano	PhD student	44
Miguel Avila Rivas	PhD student	44
Matias Medina Gonzalez	PhD student	44
Bernabe Bustos Becerra	PhD student	44
Eleodoro Riveras	PhD student	44
Pamela Naulin	PhD student	44
José Miguel Álvarez	PhD student	44
Tatiana Kraiser	PhD student	44
Viviana Araus	PhD student	44
Tomas Moyano	PhD student	44
Tomas Puelma	PhD student	44
Orlando Contreras	PhD student	44
Eva Villarroel	PhD student	44
Bernabé Bustos	PhD student	44
Calixto Domínguez	PhD student	44
Rodrigo Pulgar	PhD student	44
Calixto Domínguez	PhD student	44
Mauricio Latorre	PhD student	44
Graciela Argüello	PhD student	44
Leonardo Pavez	PhD student	44
Mariana Acuña	PhD student	44
Emilio Díaz	PhD student	44
Adriana Rojas	PhD student	44
Hugo Sepulveda	PhD student	44

Rodrigo Aguilar	PhD student	44
Fernando Bustos	PhD student	44
Claudia d'Alençon	PhD student	44
Cristian Undurraga	PhD student	44
Mario Sánchez	PhD student	44
Jorge Zúñiga	PhD student	44
Margarita Parada	PhD student	44
Joao Botelho,	PhD student	44
Gabriela Zavala	PhD student	44
Rodrigo Morales	PHD Student	44
Luis Solano	PhD student	44
Diego Rojas Benitez	PhD student	44
Consuelo Ibar	PhD student	44
Guillermo Rodríguez	Master's student	44
Daniela Elizondo	Master's student	44
Juan Pablo Parra	Master's student	44
Macarena Greve	Master's student	44
Omar Sandoval	Master's student	44
Flavia Roman Brigando	Master's student	44
Pablo Leon Medina	Master's student	44
Leandro Farias	Master's student	44
Francisco Altimiras	Master's student	44
Ricardo Gutiérrez	Master's student	44
Tatiana Opazo	Master's student	44
Daniel Meza	Master's student	44
Paulina Rudolffi	Master's student	44
María Ignacia Cadiz	Master's student	44
Kazherine Salazar	Master's student	44
José Moya	Master's student	44
Camila Mardones	Master's student	44
Oscar Peña	Master's student	44
Nicole Reynaert	Master's student	44
Consuelo Anguita	Master's student	44
Marjorie Alvarez	Master's student	44
Natalia Beiza	Master's student	44
Claudio Soto	Master's student	44
Paulina Falcón	Master's student	44
Pablo Lois	Master's student	44
Carolina Ortíz	Master's student	44
Samuel Martínez	Master's student	44
Angel Pardo	Undergraduate student	44
Alexis Peralta Carrera	Undergraduate student	44

Bernardo Pollak	Undergraduate student	44
Esteban Garate	Undergraduate student	44
José Galdames	Undergraduate student	44
Salomé Muñoz	Undergraduate student	44
Cristina Muñoz	Undergraduate student	44
Simón Carrillo	Undergraduate student	44
Daniela Ureta	Undergraduate student	44
Florencio Espinoza M	Administrator, accounting	20
Karen Meyer B	Journalist, communications	44
Carolina Oyaneder	Secretary	44
Juan Silva	Janitor and messaging	10

5. **Changes in research personnel:** Describe any changes in the principal and associate researchers relative to the original project.

   Our group of investigators remains unchanged except for the incorporation of Dr. Christian Hödar as a new Associate investigator (this was announced in last year's report). He became a full member of the center beginning in October 2013 when he concluded his postdoctoral fellowship and was hired by the Universidad de Chile.

6. **Advisory committee:** describe its tasks, the frequency of meetings, and usefulness of the advice provided to the Center. Also, report on the availability of the committee to assist the Center.

   Our International Advisory Committee did not meet with us during 2013 (they did so in December of 2012) and we have shceduled a new meeting for April 2014. They will carry out an evaluation of our performance and offer suggestions as they did previously. The results of this internal evaluation will be crucial for meeting the challenges of the final stage of the first period of funding.

III. OBJECTIVES AND RESULTS ATTAINED

1. RESULTS OBTAINED RELATIVE TO CENTER OBJECTIVES

a. Considering the objectives established in the project, in no more than 15 pages describe the results accomplished during the period. Refer also to those objectives that have not been accomplished, justifying the reasons. Organize your report describing the most significant outcomes for the following aspects:

i. Main research findings

We maintain the organization of the objectives and their respective results in the three sections described in the previous report. In short, these correspond to our defined fields of inquiry and are not coincident with the three objectives stated in the original proposal. These are, (i) genomes and regulation of gene expression in organisms living in an extreme environment, the Chilean Northern Desert (*Extreme genomes*); (ii) genome structure and functional genomics in economically or biomedically important species that are of national interest (*Relevant genomes*); (iii) the cellular and molecular basis of gene expression control for the generation of phenotypes (*Gene expression in cells*). We remind the reader that the change in objectives was been introduced as a response to the comments by anonymous peers received after our 2011 report and was approved by reviewers in the 2013 site visit (see Section V).

Research Line 1. Extreme Genomes.

a) Plants

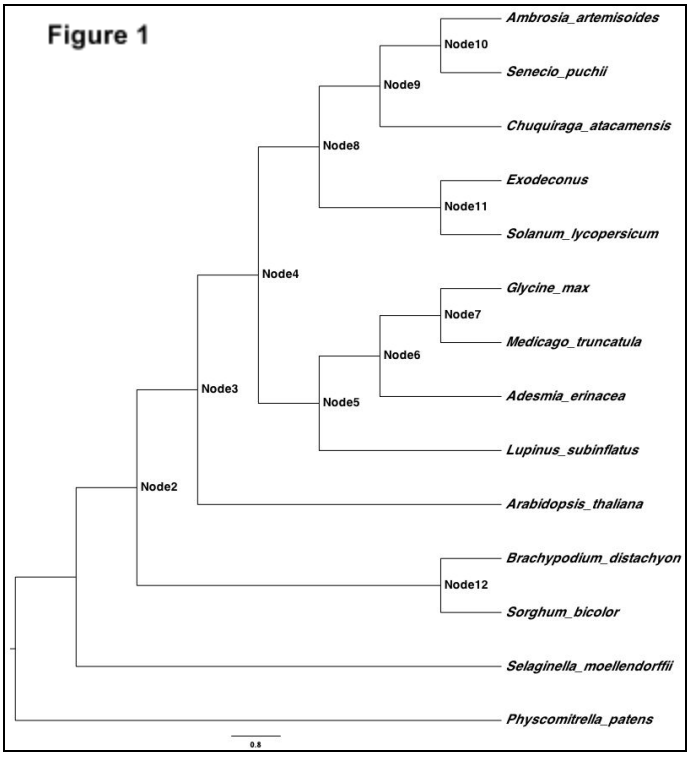
Transcriptomic analysis of plants growing on the western slopes of the Andes in the Atacama Desert reveal candidate genes underpinning adaptations to extreme abiotic stress conditions.

The Atacama Desert is one of the driest places on Earth. Plants grown here are exposed to challenging environmental conditions like extremely low water availability, large temperature oscillations, nutrient-poor soils with high salt content, and high levels of solar radiation. Despite these harsh conditions, the Atacama hosts a surprising diversity of plant life. It is poorly understood how plants can adapt to this highly demanding environment. In the central Atacama Desert, the western slopes of the Andes provide a natural altitudinal gradient of environmental conditions, such as rainfall and temperature. As a consequence, various plant communities succeed each other at different elevations: the pre-puna (2400 – 3300 m a.s.l), the puna (3300 – 4000 m a.s.l.), and the high Andean steppe (4000 – 4500 m a.s.l.). Yet, very little is known about the underlying genetic diversity of plant species inhabiting these native communities. In this study, our aim was to characterize the genetic diversity of six plant species growing on the western slopes of the Andes in the Atacama, and identify candidate genes that could underlie their extreme abiotic stress tolerance. To achieve these goals, we have chosen a transcriptomic approach, because high-throughput transcriptome sequencing has been proven to be an effective tool to reveal new molecular functions underpinning plant abiotic stress resistance. Three species belong to *Asterales*, the order represented by the highest number of species at our study site. *Ambrosia artemisioides* was collected from the pre-puna zone, *Chuquiraga atacamensis* is from the puna, and *Senecio puchii* is from the steppe. *Lupinus subinflatus* and

Adesmia erinacea are representatives of *Fabales*, they might have an important role in the nitrogen-deficient desert environment through their association with nitrogen-fixing bacteria. A desert annual, *Exodeconus integrifolius*, from the *Solanales* order, was collected in the pre-puna, and found growing even at the edge of the absolute desert. Our studies provide new insights into plant abiotic stress tolerance, and improve our understanding of the highly unique ecosystem of the Atacama Desert.

In order to understand the bases of plant adaptation to extreme environments, a phylogenomic analysis was performed on the transcriptomics data sets. This analysis was carried out in collaboration with Dr. Gloria Coruzzi's research group (New York University), using a pipeline developed for the BigPlant project of the New York Plant Genomics Consortium. First, unigenes were translated into proteins for each data set. ORFs that covered at least half of the contig were retained, the rest were discarded. These sets of peptide sequences were then aligned to a protein database of six vascular plant species and all peptides that did not have any matches were discarded. If there were still more than one valid translation for a unigene, the longest peptide was picked. Based on these aligned protein sets, a phylogenetic tree was constructed using maximum parsimony (Figure 1). After this process, the individual ortholog support was calculated for each node in the tree. Based on these data, a list of genes were created for each node, containing ortholog groups providing positive and negative Partioned Branch Support (PBS) for the given node.

Figure 1

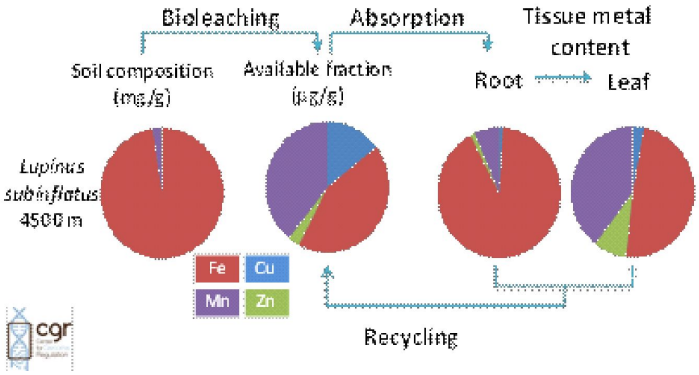


Our aim is to gain information about the adaptation of desert plants to extreme environmental conditions. Therefore, we performed a functional genetic analysis of ortholog groups providing positive support for Nodes 5, 6, and 11; these nodes represent the divergence of *Lupinus subinflatus*, *Adesmia erinacea*, and *Exodeconus integrifolius*, respectively. Ortholog groups with the highest PBS values were selected, with the top 10% chosen as a cut-off. Then, a list of *Arabidopsis* genes falling into these ortholog groups was created. Functional genetic analysis of these gene sets was carried out using the VirtualPlant platform. The BioMaps tool was used for identifying what Gene Ontology terms are enriched in each gene set, using *Arabidopsis thaliana* tair10 genome as a background population, with a p-value cut-off of 0.01. For Node 6, no significant terms were found. Node 5 and Node 11, showed several overrepresented biological processes with genes associated to nitrogen compound metabolic processes in both data sets. These results suggest an important role of nitrogen metabolism-related genes in adaptation to the extreme environment of the Atacama Desert for Hexodeconus (Solanaceae) and Lupinus (Fabaceae) species,

an ecosystem with severe nitrogen limitation. Because these species are closely related to model crops of great interest, functional studies are underway to understand the underlying mechanisms associated with N metabolism under extreme conditions in the Atacama Desert.

Chemical analysis of soil

Soil samples collected from all stations were analyzed. Elemental composition and soil texture did not show major differences along the transect indicating the soil was the same throughout it, with characteristics typical of desert ecosystems (e.g. high sand content, low organic matter). However, the soluble fraction showed marked differences along the transect. For example, we found a pronounced pH gradient, with acidic soils at high altitudes and alkaline soils at the lower end of the transect. Lower amounts of precipitation at the lower elevations correlated with higher salinity, as indicated by the elevated concentrations of Na+, K+, carbonates, and Ca2+. Stations with optimal pH for plant growth (6.6 to 7.3) hosted the highest diversity of plant species. Nitrogen, an essential macronutrient, was limiting in all soil samples. This suggests the important role of biological nitrogen fixation by plant-associated bacteria in this ecosystem. Differences of pH at different elevations showed strong correlation with the availability of several other macro- and micronutrients such as iron, phosphorus and magnesium, among others. In addition, analysis of Fe, Cu, Zn and Mn distribution among soil and plants tissues (as in *Lupinus subinflatus*, Figure 2) suggest that biolixivation, plant absorption and recycling processes determine this contrasting metals composition in the soluble fraction of the soil.



Genomics and Transcriptomics of flowering plants from the Atacama desert.

Despite the harsh conditions of the lower Atacama Desert, there are sporadically flowering plants near the coastal region and central valleys that evolved to adapt to scarce water resources and poor soil conditions. During spring a number of species flower, producing a phenomenon known as the "Blooming Desert", which takes place between 26-32°S latitude. This occurs every time the amount of rainfall observed during May to August allows the emergence of these species, conditions that are not met every year. In fact, during 2013 there was no blooming desert. However, we were able to find a few spots were *Cistanthe longiscapa* was able to grow. We collected samples from the field and RNA was prepared and transcriptome analyses of flowers, leaves (green and reds), stems (succulent and inflorescence) and roots were performed. Differential expression analyses were performed. Cluster analyses show that a number of differences are observed (Figure 3). Different patterns of gene expression can be observed in each of the organs. A more detailed analysis of the genes that are changing is currently in progress. We expect to identify genes or set of genes that show differences in expression, establishing a set of candidate genes that will be further analyzed.

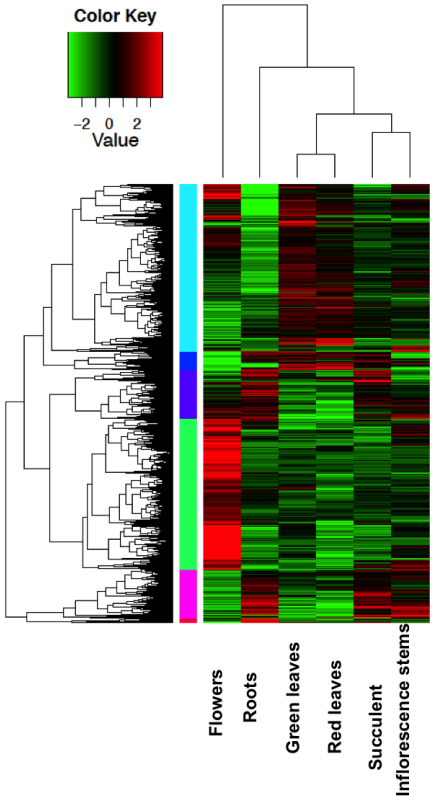


Figure 3: RNA seq analyses was performed in RNAs extracted from roots, flowers, red leaves, green leaves, suculent stems and inflorescence stems. The total number of gene products that changed in all the libraries was around 16.000.

b) Animals.
Transcriptomics of *Rhinella Spinulosa*, an Andean amphibian with a unique life history.

We have completed the first phase of this study where we identify -through transcriptomic analysis- genes potentially responsible for local adaptation in the Andean frog *Rhinella spinulosa*. We refer the reader to the Appendix, where we supply a submitted manuscript describing the results of this project (Pastenes et al., 2014). Briefly, this species has a wide distribution along the entire mountain range of Chile, living at altitudes of 1500 to 4500 meters above sea level. In particular, we are interested in two populations of this species that live in divergent environments. There is one population, representative of the typical environment that these animals inhabit, living in the mountains of the Central region (near Santiago), where there are clearly defined seasonal changes in temperature and water availability. In this environment, the animals essentially hibernate for the winter (when there is abundant snowfall) and reproduce in spring. Embryos and larvae develop fast and metamorphose rapidly to avoid dissecation of the ponds they inhabit in summer. These animals are very tolerant of wide temperature variations and can be bred and grown in captivity at 20ºC. On the other hand, there is a population in the Tatio geysers, about 1500Km north of Santiago and at an altitude of about 4000 meters. Here, the frogs live in streams that have a constant temperature of 25ºC, reproduce all year

long and development proceeds at a slower rate than in the first population; metamorphosis occurs later as well. Despite being populations of the same species, there have obviously been physiological, reproductive and behavioral local adaptations that merit study. As an experimental design, we have generated conditions of growth in artificial environments for the animals, and we have reared them at two temperatures. We have included animals from a control population as well, and two developmental stages have been sampled. The twelve contrasting RNA samples have been analyzed by next generation sequencing (NGS) and the transcriptomes were been obtained. Full-length mRNAs were assembled and a total of about 30,000 transcripts were found. Since this is not a model organism, and there are very few sequenced amphibian genomes, we had to use *de novo* annotation of the sequences and assign orthology using our bioinformatics platform at the CGR. Among the genes identified as being differentially expressed in the different populations and different environmental regimes, we found several peptide hormones known to be involved in metamorphosis and transcription factors that mediate the stress response (Pastenes et al., 2014).

Genome evolution of fish of the genus *Orestias*, a killifish that inhabits the high altitude salt lakes.

Killifish are among the most widely adapted and dispersed group of teleosts (bony fish) and those of the genus *Orestias* that live in the salt lakes of the *Altiplano* display a remarakble evolutionary history. They have speciated in single salt lakes along the Andes in allopatric fashion and have adapted to different environmental conditions, most notably, salinity of the water they inhabit. As a prime example, *Orestias ascotanensis* was dentified as one of the species that we wanted to characterize at the genome level since the beginning of our Center. To date, we have a first draft of the complete genomic sequence of *O. ascotanensis*. The data we have obtained from the genomic DNA prepared from a single individual is outstanding and shows that this species, unlike other teleosts sequenced thus far, has a genome that is relatively free of repetitive elements, which normally make assembly exceedingly difficult (i.e., Atlantic salmon; see below). A summary of the data can be seen in the table shown here. Genomic DNA of a single male *O. ascotanensis* was sequenced and assembled to a size of 0.695 gigabases, comprising 2,394 scaffolds (N50 contig, 43.5 kilobases; N50 scaffold, 2.66 megabases; see Table).

Atrribute	*Value*
Contig minimum size for reporting	1000
Number of contigs	30,257
Total contig length	670,073,028
N50 contig size in kb	43.5
Max Contig length	365,805
Number of scaffolds	2,394
Total scaffold length, with gaps	695,674,091

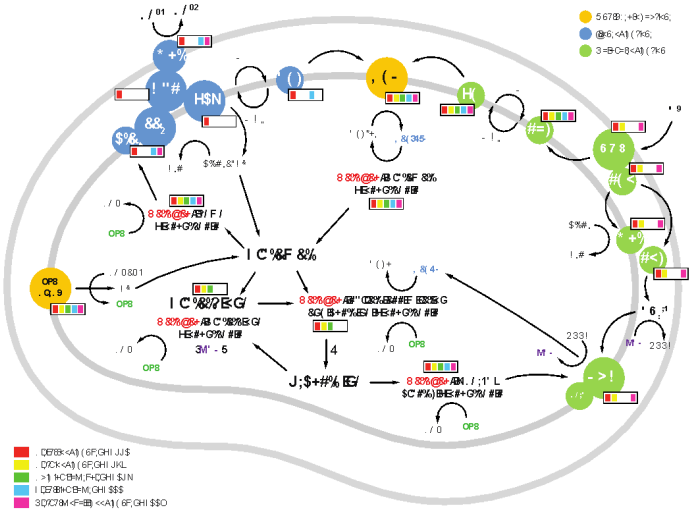
N50 scaffold size in kb	2,666
Max Scaffold lentgth	14,017,584
Median size of gaps in scaffolds	304
Median dev of gaps in scaffolds	61
% of bases in captured gaps	3.71

c) Microbes.

Systems biology for biomining bacteria

The purpose of this project is to unravel the main biological mechanisms explaining the capacity of the principal bacterial consortiums isolated from copper mines in the north of Chile to bioleach copper mineral and to live in extreme environments. We adopted a systems biology approach trying to integrate differents layers of omic information. During 2012 we sequenced draft genomes using NGS technologies for the 6 main indigenous biomining bacteria: *A. ferrooxidans, A. thiooxidans, A. cryptum, L. ferriphilum, S. thermosulfidooxidans,* and *Ferroplasma sp.* The draft genome and annotation of *Sulfobacillus thermosulfidooxidans* was already published in J. of Bacteriology during 2012. A comparative analysis based on the distance of ortholog genes confirmed the classification of such species from their complete genomes. During this year we have worked on the characterization of resistence mechanisms to heavy metals of this consortium. During the bioleaching process, the increase of soluble oxidized metals can induce the formation of reactive oxygen species (ROS). Although, all the bacteria contain the classic ROS response genes (SOD, catalases, peroxidases, etc.), the number of these components does not differ from others organisms, suggesting that the high resistance to the environmental possibly lies in its ability to handle the concentration of metals, avoiding its intracellular toxic effects. In this context, it was not surprising the finding of high amount of metal resistance determinants, principally the existence of ATPases Type P and chaperons. In particular, the elevated number and duplication of copper proteins confirms the point that intracellular copper unlike zinc, arsenic and iron is the most toxic element for the consortium. Interestingly, *Acidiphilium sp* compared to others heterotroph presents an elevated number of copper resistance proteins with a high number of transcriptional factor regulator, suggesting mayor complexity in this bacteria related with the others species of the consortium. For the case of the gram-positive *S. thermosulfidooxidans*, the lack of an outer membrane prevents the existence of RND metal proteins, to compensate this absence, a high number of CDF and principally CopA copper resistance proteins are codified in the bacterium. Regarding iron resistance mechanisms, *A. ferrooxidans* encodes a higher number of proteins involved in iron handling, including principally permeases and iron transport systems. In terms of the gene regulation of these proteins, all bacterial species encode at least for one Fur family members (the principal iron uptake transcriptional regulator). In particular, *S. thermosulfidooxidans* and *L. ferriphilum* have three copies of this gene. A manuscript is in the last stage of preparation for submission involving this data.

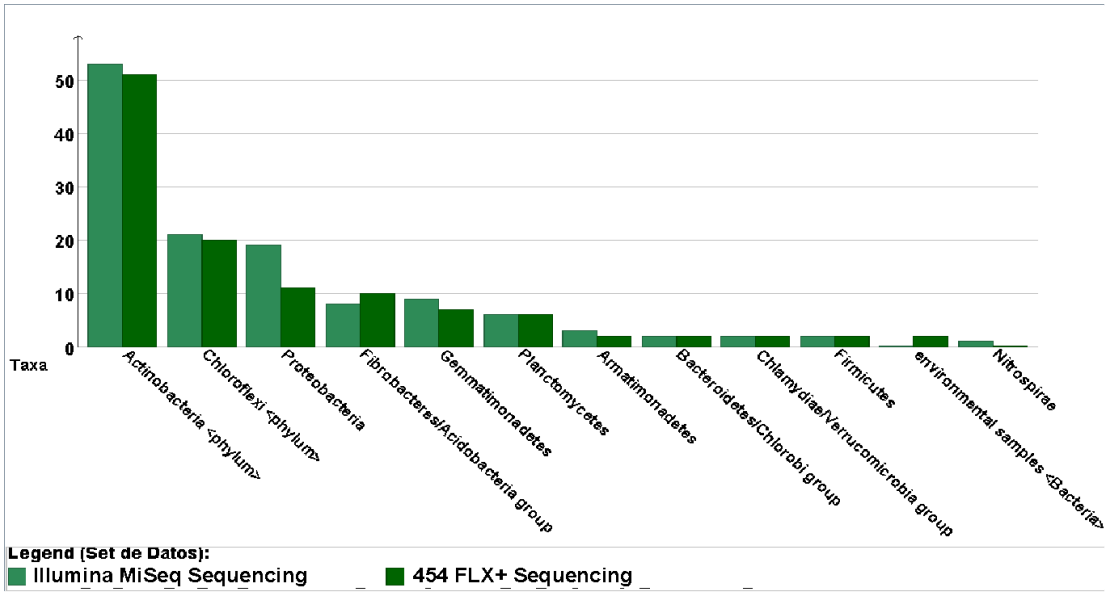
Figure 4. An integrative consortium metabolic pathways model, capable of assigning species specific components, into specific metabolic routes.



In complementary work, during this year we finished the computation modeling of the 3d structure protein Licanantase (Lic), which is the major component of the secretome of *Acidithiobacillus thiooxidans* when is grown in elemental sulphur. When used as an additive, Lic is able to improve copper recovery from bioleaching processes. The idea is to obtain insights on its structure-function relationships and to shed light on its role during the bioleaching process. An important conclusion of this work was that Lic stability can be explained by a higher number of both intermonomer hydrophobic contacts and hydrogen bonds, which would maintain the secondary structure. Thus, Lic would be stabilized by pH-insensitive nonbonding interactions, which would provide it acid resistance at the extreme low pH where it performs its function.
The work developed and reported during 2012 about the metabolic modeling of prokaryotic oxidation of reduced inorganic sulfur compounds (RISCs) appeared this year in the journal Biotechnology and Bioengineering (Bobadilla-Fazzini et al., 2013). Additionally, the first metabolic profiles for two bioleaching bacteria using CE-MS. *A. ferrooxidans* and *A. thiooxidans* appeared in the journal Metabolomics (Martínez et al., Metabolomics 9:247-257. 2013).

Comparative metagenomics of bacteria across an altitude gradient in the Atacama Desert.
With the aim to explore all ecological variables that could contribute to facilitation and plant survival in the previously described altitude gradient in the Atacama Desert, we decide to study microorganism composition of these soils looking for a correlation between they distribution and richness and the contrasting composition in the soluble fraction of the soil. It's known that microorganisms can contribute to nutrient bioavailability and/or pathogen displacement and that could establish an ecological link, especially in extreme environments. We extracted DNA from soil close to the location of sequenced plant to characterize on site microorganisms using a metagenomic approach. With this objective we are currently testing different DNA sequencing technologies trying to define the one that fits better to our requirement. For these experiments we are using DNA extracted form site 12 that is where we have obtained more DNA per gram of soil. We are testing the following technologies: 454 FLX plus, Illumina MiSeq250 and Illumina HiSeq2000 (one plate each). Currently we have results for

the first two techniques with 1.000.000 and 20.000.000 reads respectively with 700Mb and 5Gb of total data respectively. When we search for 16S ribosomal gene represented on each sample and its abundance, the results are comparable, with 50% of r16S genes belonging to Actinobateria phylum, 20% to Chloroflexi and Proteobacteria, almost 10% to Fibrobacter and Gemmatimonadetes, 5% to Planctomycetes and other 5 phylum that have less than 3% of representation. **Figure 5** below shows bacteria abundance (by phylum) identified at 4,470msnm at the Atacama Desert.



Research Line 2. Relevant Genomes.

a) *Homo sapiens*, the Chilean Human Genome Project (the Mapuche-Huilliche population)

This objective has culminated in the submission of a manuscript to the journal Nature (under review as of this writing). The following is an excerpt from the manuscript, which is also included in the Appendix (Vidal et al., 2014). Please refer to the figures within the manuscript.

The Southern Amerindian Genome Reveals Links to Prevalent Diseases in Native and Admixed Latin Americans: Sequencing the complete genome of 11 individuals belonging to a native Mapuche population, from the south of Chile.

Current high-coverage full genome efforts have mostly focused on Old World continental groups (Europeans, Asians and Africans) and there is scant information concerning the genetic structure of ancestral American groups. Contemporary Latin Americans, 400 million of which live in South America today, are predominantly the result of admixture between Native Americans and Europeans. Therefore, addressing genetic variation in Native populations should help us understand the molecular basis for common diseases affecting modern Native and admixed Latin American populations.

The Mapuche is considered to be one of the largest original populations of the South Cone. We selected individuals from the Mapuche-Huilliche ethnic group from a previous health survey in the

south Chile to ensure as little admixture as possible. Most of the individuals selected have gallstone disease and some of them other common traits among this population and in Chile such as obesity, insulin resistance, type 2 diabetes (T2D) and hypertension. DNA samples were sequenced using the combinatorial probe-anchor ligation and DNA nanoarray technology of Complete Genomics. Sequences obtained corresponded to at least 80% of the entire genome and 98% of exonic regions of each individual with at least 30X coverage and 96-97% high-confidence calls.
A genome-wide summary including main classes of genetic variants in Mapuche-Huilliches is presented in Figure 1 (Vidal et al., 2014). Approximately 3.1 x 10^6 high-quality single nucleotide variants (SNVs) were determined for each individual, totaling 5,853,203 SNVs in the cohort. We find a high level of agreement (99.70%) in the SNV calling rate between genome sequencing and SNV genotyping results when using the Illumina Infinium HumanCoreExome BeadChip as a validation control. We also identified 464,952 (7.9%) novel SNVs not included in the latest release of dbSNP build 138 or not having a reported frequency in the 1,000 Genomes Project phase 1 database. Likewise, analysis of small-scale genetic variants indicated that 270,640 (66.5%) insertions and 61,780 (14.6%) deletions are novel and observed in at least 1 Mapuche-Huilliche genome. Analysis of large-scale genomic events detected 680 copy number variants (CNVs), including 19 novel events not reported in the Database of Genomic Variants (7). We also found 4,515 structural variants (SVs) out of which 350 are considered novel. A high proportion of common events affecting at least 4 individuals were detected: 35% and 21% CNVs and SVs, respectively. While 1,102 genes were found partial or completely overlapped by these types of genomic variants in at least one individual, there were 22 protein-coding genes consistently affected in all Mapuche-Huilliche genomes (fig. S2). Interestingly, we found a novel CNV event affecting exon 2 of the dimethylarginine dimethylaminohydrolase 1 (*DDAH1*) gene. The product of this gene regulates cellular levels of methylarginines, is involved in nitric oxide generation and may be important for early circulatory dysfunction and coronary artery disease. We also observed two large exonic deletions within the complement component (3b/4b) receptor 1 (*CR1*) gene. Expression levels of this protein and/or mutations in its gene have been associated with risk to develop gallbladder carcinoma, a fatal complication of gallstone disease, prevalent in Native and admixed Americans and in the Mapuche-Huilliche cohort. These results indicate there are multiple genomic regions enriched with small- as well as large-scale genetic variants and highlight zones with strong variability that are private to Mapuche-Huilliche individuals.

Sequenced genomes represent an original non-admixed American population that could serve as a high quality reference for Native Americans, as well as populations with Amerindian ancestry.

The Mapuche are the modern representatives of one of the most prominent indigenous groups in the Southern Cone of South America. Mapuche people descend from early hunter-gatherers who colonized the subcontinent about 15,000 years ago. Several decades of research suggest that most Native Americans descend from a single first ancestral population flow from Asia, known as 'First Americans'. Peopling of Central and South America is thought to derive from these 'First Americans', following a southward expansion along the coasts eventually reaching Patagonia. To gain insight into the ancestry of this selected group of Mapuche-Huilliches, we determined their maternal lineage by analysis of mitochondrial DNA genomes. We found that all 11 individuals belong to the Native American haplogroups C and D, two of the founder major pan-continental haplogroups. The majority of Mapuche-Huilliches sequenced (7 out of 11) belonged to the C1b haplogroup and 6 of them were assigned to the clade C1b13, which is a branch found mainly in the Southern Cone of South America

between 38° and 42°S. While the other 4 individuals belonged to the D haplogroup, three of them are in the D1g clade, which is found almost exclusively in the central-southern part of Chile and Argentina, and only one was in the D4h3a clade, found mainly in southern Patagonia. Thus, our cohort shares maternal lineages with other contemporary native populations from central and southern Chile and Argentina, indicating they are representative of Native Americans inhabiting this part of the continent.

To determine the degree of non-native American admixture of our genome samples, we compared their genomes to those of Yorubas from Ibadan, Africa (YRI), Chinese Han from Beijing (CHB) and Utah residents with European ancestry (CEU), all of which represent ancestral founders. The result of ADMIXTURE analysis is consistent with four population groups (K=4) (Figure 2, Vidal et al., 2014). Notably, in our ADMIXTURE model, most of the genetic contribution in the Mapuche-Huilliche genomes comes from their own Amerindian ancestry, with negligible contributions of Asian, European and African ancestries. This result is in agreement with the idea that these Mapuche-Huilliche individuals correspond to an isolated Native South American population. Consistently, when we include in our analysis sequence data from an American Admixed population from Los Angeles, USA, with Mexican Ancestry (MXL) (Figure 2), the Mapuche-Huilliche population behaves as a founder population for admixed Mexicans, contributing with approximately 40% of their genetic composition. To further explore the structure of the cohort, we performed a principal component analysis (PCA) using EIGENSTRAT. We used 474,569 SNVs that are present in at least 1 Mapuche-Huilliche individual and shared with at least one individual from a diversity panel that includes Africans, East and South Asians, Europeans and Admixed Americans (Figure 2). Mapuche-Huilliches, as well as African, European and Asian populations, form distant and defined clusters (Figure 2), in agreement with a lack of recent admixture. We also find that Mapuche-Huilliches are closer to the East Asian cluster than to the European, South Asian or African clusters, in accordance with the accepted migrational model. Admixed American populations scatter between the clusters of European and Mapuche-Huilliche populations, as expected for Americans having diverse non-native genetic contributions. Finally, phylogenetic analyses using MrBayes software reveal that Mapuche-Huilliches are located near the East Asian clade, as shown by the PCA analysis (Figure 2C), and that they cluster together in a separate branch located within the Admixed American clade (Figure 2D). These results indicate Mapuche-Huilliches represent an original non-admixed American population that is clearly distinguished from other ethnic groups. Therefore, genetic variations in Mapuche-Huilliches represent a key source of data for identifying markers that are specific for American Amerindians and Mestizo populations with Amerindian ancestry.

Genome sequence information provides insights to explain susceptibility to common complex diseases in Mapuche-Huilliches.

It has been shown that Native American and Mestizo populations with Amerindian ancestry are more susceptible than other populations to develop certain diseases, most of them related to metabolic disorders, especially when they adopt modern urbanized lifestyles, including gallbladder disease (GSD), T2D, obesity and insulin resistance. Native populations of North and South America share traits and susceptibility to develop these and other chronic metabolic disorders which are partially explained by specific genetic factors that are frequent in and/or private of Native Americans. To identify SNVs with a potential functional impact that could be associated to these diseases/traits in our cohort we first used the GWAS catalog to determine SNVs that have been associated to human diseases or traits and that have a significantly higher allelic frequency in Mapuche-Huilliches as

compared with world populations included in the 1,000 Genomes Project phase 1 and shared by 3 or more individuals. We found 401 variants, and 432 related genes from this analysis. Interestingly, we found 56 variants directly associated with common diseases present in the Mapuche-Huilliche cohort that may contribute to the higher prevalence of these diseases in this population. We next used SNPEff to determine SNVs located in protein coding genes that were classified with a "high" or "moderate" impact. From these, we selected 4,405 variants that had a significantly higher allelic frequency in Mapuche-Huilliches, and that were shared between 3 or more individuals. 1,313 of these variants, corresponding to 1,124 genes, had a potential functional impact as determined by dbNSFP analysis. To complement our analysis, we also included Indels that had a "high" or "moderate" impact according to SNPEff, for a total of 1,271 genes. In sum, using two complementary approaches, GWAS information and functional impact prediction, we were able to find 1,636 genes with variants that may have a potential impact on gene function.

Consistent with the phenotypes of the individuals, we found an overrepresentation of genes related to diseases and traits present in the Mapuche-Huilliche cohort according to Disease Ontology analysis: *Diabetes mellitus* (58 related genes), hypertension (23 genes), obesity (23 genes), cholelithiasis (8 genes), hyperlipidemia (7 genes), hypercholesterolemia (7 genes). Additionally, we found an overrepresentation of genes in KEGG pathways (*31*) related to lipid and sugar metabolism and chronic inflammatory diseases.

In order to determine functional interactions between the list of genes with variants with potential functional impact, we constructed a network graph using the 1,636 genes obtained above from GWAS and functional analysis as nodes and molecular interaction data as edges. Additional edges were drawn between genes and diseases/traits from the GWAS catalog, also represented as nodes in the network. We were able to find evidence to connect a total of 581 nodes and 900 edges (Figure 3). To identify functional subnetworks, we used the Antipole network graph-clustering algorithm, and functionally classified the resulting clusters by either diseases/traits assigned to selected genes, or by overrepresented biological processes of the genes contained in the clusters (Figure 3). We found 559 genes clustered in nine different subnetwork modules. It is noteworthy that the largest disease-related module contains genes related to metabolic and/or immune system disease (Figure 3). It has been shown that the immune system and metabolism are highly integrated processes. Particularly, inflammation associated to obesity is one of the main causes of insulin resistance and diabetes. Accumulation of specific variants with a potential functional impact in highly connected metabolic as well as immune system disease-related genes, along with alteration of other genes in interconnected clusters, may contribute in part to the phenotype of Mapuche-Huilliches. Consistently, 20 out of 48 genes associated with T2D or glucose/insulin levels (*RASGRF, G6PC2, PDGFRA, SLC30A8, GCK, HNF1B, PRC1, FADS1, GCKR, IRS1, KCNQ1, MADD, ADAMTS9, CDKAL1, CDKN2A, CDKN2B, HNF1A, IGF1, KLF14, PPARG*) and 13 genes found in loci associated with obesity (*LEPR, TMEM18, BDNF, SH2B1 MTCH2, CDKAL1, SEC16B, TFAP2B, RABEP2, ZNF608, GP2 and KLF9)* were found in our network. Moreover, a recent report identified a risk haplotype for T2D with a high frequency in Native and admixed Americans, consisting of five SNVs in the *SLC16A11* gene. We found that these SNVs had a significantly higher allelic frequency in the Mapuche-Huilliche cohort, as compared with 1,000 Genomes phase 1 populations, along with 21 other SNVs that had been associated with T2D in different studies, suggesting a strong burden of genetic risk for this common metabolic condition in this cohort of Native Americans.

b) Vitis vinifera (Thompson seedless or 'Sultanina') genome.

Grapevine (*Vitis vinifera* L.) is the most important Mediterranean fruit crop, used to produce both wine and spirits as well as table grape and raisins. Wine and table grape cultivars represent two divergent germplasm pools with different origins and domestication history, as well as differential characteristics for berry size, cluster architecture and berry chemical profile, among others. 'Sultanina' plays a pivotal role in modern table grape breeding providing the main source of seedlessness. This cultivar is also one of the most planted for fresh consumption and raisins production. Given its importance, we sequenced it and implemented a novel strategy for the *de novo* assembly of its highly heterozygous genome. Our approach produced a draft genome of 466 Mb, recovering 82% of the genes present in the grapevine reference genome; in addition, we identified 240 novel genes. A large number of structural variants and SNPs were identified. This year we proceeded to validate some of them: 45 (21 SNPs and 24 INDELs) were experimentally confirmed in 'Sultanina' and six SNPs in other 23 table grape varieties. This work produced the first structural variants and SNPs catalog for grapevine, constituting a novel and very powerful tool for genomic studies in this key fruit crop, particularly useful to support marker-assisted breeding of table grapes. The article was accepted in BMC Plant Biology (DiGenova et al., 2014).

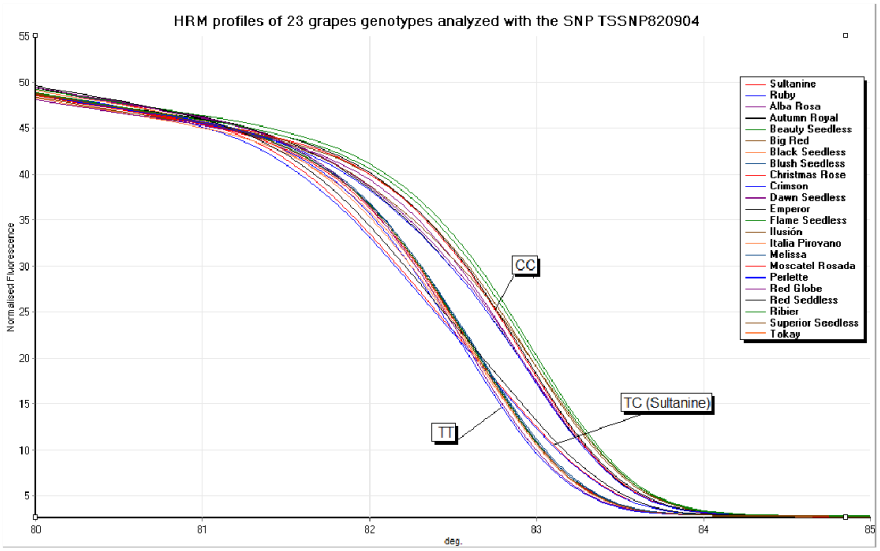


Figure 6. HRM profiles of 23 table grape varieties for the SNP TSSNP820904.

The HRM analysis produced robust results confirming the transferability of the SNP TSSNP820904 (T- >C) in varieties with different genetic background.
In the same genome, a first result of a large scale mRNAseq experiment was the proposition of the grapevine genes *VvAIG1* or *VvTCPB* or both as a reference tool to normalize RNA expression in qPCR assays or other quantitative method intended to measure gene expression in berries and other tissues of this fruit crop, sampled at different developmental stages and physiological conditions. During 2014 we expect to publish the complete analysis of this experiment.

c) *Salmo salar*, the Atlantic salmon genome

We have continued collaborating very closely with the International Consortium for Sequencing the Atlantic Salmon Genome (ICSASG), verifying data and producing our own computations and assemblies. In particular, during 2013 we have produced the All-path assembly of the salmon genome

that have been used in the reconciliation of the sequence to be declared the first draft of the salmon genome in june 2014 reaching the objectives of the consortium: to produce the best possible reference Atlantic salmon genome and the highest possible impact article for the Atlantic salmon genome sequence. Our contribution as well as all the results of this large scale sequencing effort will be submitted for publication to a high impact journal during the first semester of 2014.
In addition, we have finished and submitted the article of a new bioinformatic pipeline for variant search and classification (SnpSACK), through an assessment of local conservation levels over the entire genome, starting from a highly homozygous reference genome and genomic sequences from multiple individuals, allowing the selection of highly conserved SNPs a priori, thus reducing the number of PSVs present in the final SNP set. The classification strategy relies on a fuzzy logic approach. SnpSACK protocol was benchmarked against Atlantic salmon (Salmo salar sp.) public data, classifying 13K SNPs as highly conserved (starting from 18K variants); also declaring as conserved 1.5K short Deletion/Insertion Polymorphisms (DIPs). The 13K SNPs obtained marked 13% of the Atlantic salmon genome. The conserved regions proposed by the SnpSACK pipeline were verified against a 5K set of conserved variants empirically validated and classified, showing that 92% of validated variants were located within proposed highly conserved regions. SnpSACK identifies PSV enriched regions, allowing the exclusion of PSVs while effectively retaining great part of the highly-conserved SNPs. Through the assay of SnpSACK against highly-duplicated species, it has shown to be an effective bioinformatic tool for the discovery, filtering, and classification of SNPs, while excluding poorly conserved variants from the final set. This pipeline is potentially applicable to any organism with a reference genome, as well as multiple sequencing experiments. Moreover, the database obtained from the pipeline contains additional information such as context, nature and source of the variant, which can be used for further studies.

d) *Piscirickettsia salmonis* (an aquaculture pathogenic bacterium)
d.i) Selenium, selenoproteins and resistance to *Piscirickettsia salmonis* infection in Atlantic salmon macrophages.
Selenium (Se) is an essential micronutrient required for the optimal functioning of the immune and redox system. We evaluated the effect of Se on the survival and anti-oxidative response of Atlantic salmon macrophages (SHK-1 cell line) infected with *Piscirickettsia salmonis*. Maximum concentration and time of exposure to selenium in SHK-1 cells were determined (no effects on cell viability relative to the basal culture conditions). These conditions were used to compare cell viability and oxidative status between uninfected and *P. salmonis* infected macrophages. Macrophages supplemented with Se showed a better survival and a lower oxidative status than non-supplemented macrophages in response to infection (Figure 6), implying a protective effect of Se, possibly exerted by the increased activity of selenoproteins, which are a diverse group of proteins that contain selenium (Se) in the form of the amino acid selenocysteine (Sec). Sec is encoded by the STOP codon (UGA). The presence of an RNA stem-loop structure, which is the Sec Insertion Sequence (SECIS) element in the 3' untranslated region (UTR) of eukaryotic mRNAs, differentiates the Sec or STOP function of UGA codons. Because of the dual meaning of the UGA codon, selenoprotein genes are often miss-predicted by standard annotation pipelines. Here, we applied a series of bioinformatics tools to predict the Atlantic salmon selenotranscriptome from large collections of EST sequences. This protocol allowed the prediction of the structure of Atlantic salmon SECIS and the identification of a set of 31 transcripts, representatives of 18 families of Selenoproteins (Figure 7).

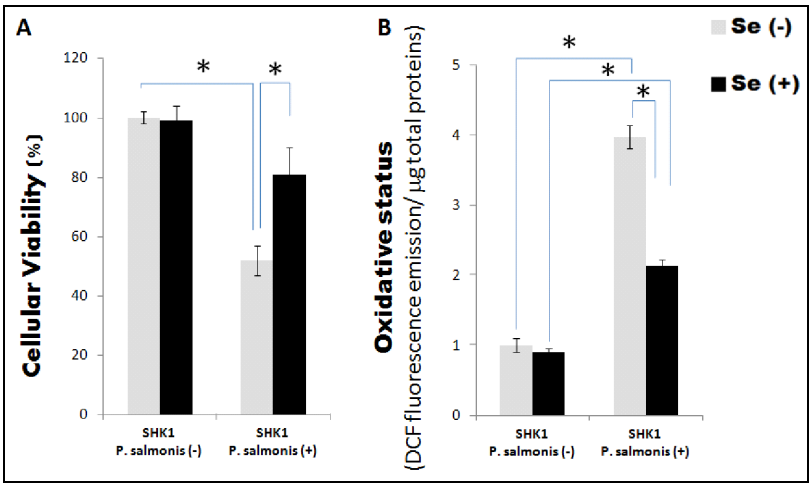


Figure 7. Cellular viability and oxidative status of salmon macrophages infected with *P. salmonis* and exposed or not to selenium. (A) Cells were cultured for 48 hours in the presence (Se +) or absence of (Se-) 1 mM sodium selenite and then infected *in vitro* with *P. salmonis* (P. salmonis (+)) or maintained in basal medium (P. salmonis (-)) for the next 72 h. Bars represent the average of twelve measurements. (B) Estimation of oxidative status was conducted by ROS measurements by quantifying the fluorescence emitted by dichlorofluorescein diacetate (DCF). Results were normalized by the total protein mass and each bar represents the average of twelve measurements. (*) significant differences between conditions (P <0.05 Student t test).

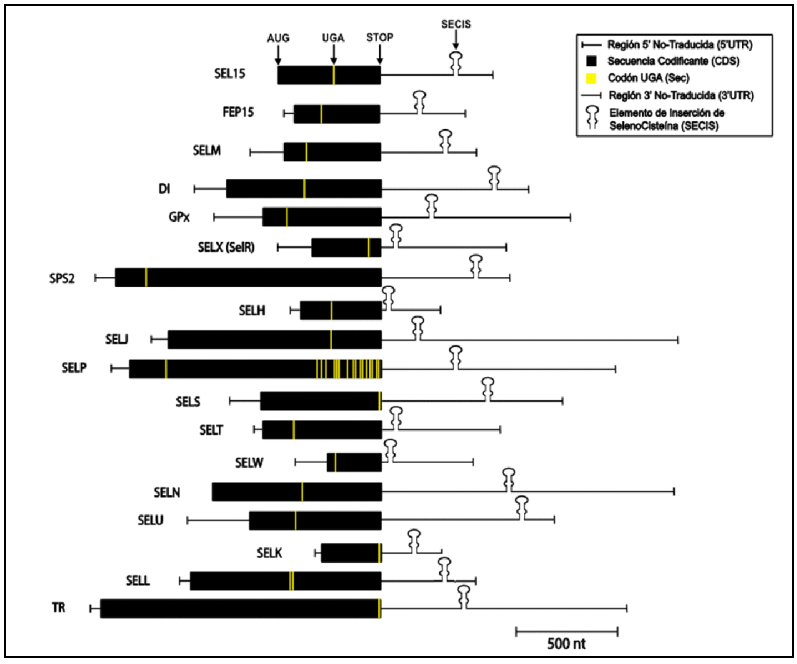


Figure 8. The Atlantic salmon selenotranscriptome. The selenotranscriptome is composed by 18 transcript families (31 individual mRNAs) with UGA in-frame codon and Selenocysteine Insertion Sequence Element (SECIS) in the 3' untranslated region (3'UTR). The figure shows the UGA in-frame position within the CDS, the relative length of the selenotranscripts and the relative SECIS distance to the STOP codon.

d.ii. Expression of *P. salmonis* iron acquisition genes during growth with different iron availabilities.

When challenged with limiting iron concentrations such as those encountered in host tissues, pathogenic bacteria handle iron homeostasis by increasing the expression of iron acquisition systems. This strategy allows bacterial survival, since iron is an essential cofactor for many proteins mediating electron transfer and redox reactions. Here we report the analysis of transcript levels for *P. salmonis* genes involved in iron

acquisition when this bacterium subjected to iron-deficit and iron-excess conditions in order to better characterize potential virulence determinants of *P. salmonis*. The *P. salmonis* genome encodes at least 44 CDS that are related to iron acquisition according to our annotation. As expected, several of these CDS presented changes in their transcript levels in response to iron availability (Table 2). For instance, after 5 days under conditions of iron deficiency (0 mg/L) we detected, by RT-qPCR, increases in transcript levels for: receptors of Fe(II) (feoB) and siderophores (fhuABCD), genes involved in the synthesis (PvsADE) and transport (pvuA) of siderophores and members of the TonB-ExbBD energy-transducing system (Table 2). Thus, in this bacterium as in other, the lack of iron results in the derepression of an entire collection of genes for the biosynthesis and transport of siderophores and hence the activity of one or more high-affinity iron uptake systems.

Table 2: Expression of *P. salmonis* iron acquisition genes

Relative expression*

1. Synthesis and transport of sidephores

Suplemento de Fe (III)	PvuA	PvsA	PvsB	PvsC	PvsD	PvsE
0 mg/L	23.21	131.5	27.44	94.49	62.55	100.35
10 mg/L	2.43	21.91	27.22	5.26	26.31	0.82
100 mg/L	3.27	15.11	22.99	1.04	11.37	65.5

2. Siderophore, heme and Fe(II) transporters

Suplemento de Fe (III)	FhuA/heme	FhuB	FhuC	FhuD	FeoA	FeoB
0 mg/L	37.51	177.06	136.52	38.02	22.24	10.48
10 mg/L	0.14	8.85	2.78	4.05	28.13	0.03
100 mg/L	0.03	7.9	30.18	7.73	707.81	0.04

3. Generation of energy

Suplemento de Fe (III)	TonB	ExbB	ExbD
0 mg/L	143.16	226.38	86.67
10 mg/L	17.22	15.32	9.47
100 mg/L	19.06	15.6	25.25

e) *Prunus persica*. The peach genome.

The paper published by the international peach genome initiative, consortium to which we belong, was published in Nature Genetics (IPGI, Nature Genetics 37: 549–554 (2013)). This is a high quality draft sequence that will expedite all future omics work performed in the *Rosaceae* family.

The proteome of peaches during ripening.

In order to identify molecular components associated to fruit quality during peach ripening we performed a large-scale peach proteome analysis using SDS-PAGE shotgun proteomics. This new approach allowed the identification of about 4,000 proteins that were mapped in metabolic and signaling pathways giving a high coverage of the peach fruit proteome. In addition, proteins belonging

to different biochemical pathways and networks, important for peach fruit ripening, were identified. Around 100 proteins related to sweetness, aroma color and texture were differentially expressed between mature and ripe fruit. These results are an important step to understand the molecular insights of peach ripening.

The transcriptome of peaches during postharvest. (Collaboration with Reinaldo Campos, UNAB). The peach chilean industry exports most of its fruit production, which requires cold storage for long distance shipments. Unfortunately, this procedure leads to a disorder known as chilling injury. Different varieties are affected in a different manner. In addition, different postharvest treatments such as controlled atmosphere, temperature conditioning, and inhibitors of ethylene have been utilized to decrease the damage produced by chilling injury, which seriously affects the peach industry. We found that early season varieties are less prone to become mealy in comparison to late varieties. On the other hand, the mealiness of one variety can be prevented using controlled atmosphere (CA) whereas the mealiness of a different variety does not respond to CA but it responds to conditioning. Transcriptome analyses of all the varieties and posthaverst treatments were performed using the Illumina platform. The analyses of this data is currently is in progress and we expect to identify genes that are related to mealiness.

f) Integrating the Potato Genome with Genetic and Physical Maps: The genome of the potato (*Solanum tuberosum*), a major global food crop, was recently sequenced. We have participated in the consortium responsible of this work. Now, we have participated presented in the integration of the potato reference genome (DM) with a new sequence-tagged site marker-based linkage map and other physical and genetic maps of potato and the closely related species tomato. This work has led to a greatly improved ordering of the potato reference genome superscaffolds into chromosomal "pseudomolecules".

Research Line 3. Gene expression in cells.

a) Epigenetic mechanisms

The epigenetic mechanisms that control the expression of lineage-specific genes during both mesenchymal and neuronal differentiation.

Among the principal epigenetic regulators described are the Polycomb-Group proteins (PcG). These proteins form complexes that are recruited to target genes where they can either favor silencing or activation of transcription during differentiation. In collaboration with Dr. van Zundert at UNAB, we have analyzed the expression of PcG proteins Ezh1 and Ezh2 (catalytic subunits) during hyppocampal development and explore their specific contribution during transcriptional regulation of genes that are critical for neuronal plasticity (Bustos et al. 2013). We find that both proteins are initially expressed in neurons and astrocytes, although their expression levels depend on the developmental stage analyzed: Ezh2 is expressed high in neuroprogenitors to then rapidly decline along differentiation, whereas Ezh1 expression persists with neuronal maturation. In parallel, a switch in their association with target promoters, from Ezh2 to Ezh1, can result in either gene activation (neuronal genes associated with plasticity, see Figure 8) or silencing (e. g. genes associated with mesenchymal lineages). This exchange occurs concomitantly with the presence of a specific pattern of epigenetic marks at histones and additional epigenetic regulators associated with promoter sequences that surround the transcriptional start site. We are currently exploring the molecular mechanisms that support the role of this Ezh2/Ezh1 switch either activating neuronal plasticity genes or repressing non-neuronal genes.

Additionally, in collaboration with Dr. Gutierrez (CGR) we are defining the role of specific miRNAs (miRs) in controlling the down-regulation of Ezh2 gene expression (as well as of other epigenetic regulators shown in the figure below). Dr. Laura Guajardo (post-doctoral fellow) has used microarray and sequencing analyses, to define a reduced population of miRs that are differentially expressed during maturation of hypoccampal neurons and that are predicted to interact with sequences at the 3'-end of the Ezh2 mRNA that are evolutionary conserved. We are currently in the process to demonstrate whether some of these miRs regulate Ezh2 expression in these cells, using both *in vivo* and *in vitro* approaches.

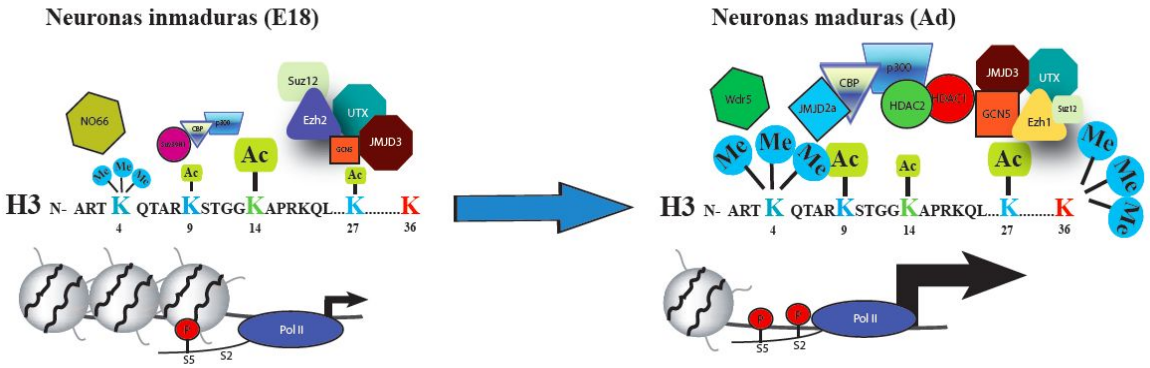


Figure 9. Epigenetic control of genes during maturation of hipoccampal neurons. Schematic representation of the protein components mediating the epigenetic control of transcription of genes that regulate neuronal plasticity in the hippocampus. Polycomb Group (PcG) complex components (e. g. Ezh1 and Ezh2: histone H3 lysine 27 methylases) and other associated activities (e. g. HDACs: Histone deacetylases 1, 2 and 4; Suv39H1: histone H3 lysine 9 methylases) are enriched at the promoter regions of the PSD95 gene during transcriptional repression. COMPASS complex components (e. g. Wdr5: histone H3 lysine 4 methyltransferase associated protein) and associated proteins (e. g. p300: histone H3 acetyl transferase), together with RNA polymerase II become highly enriched at these promoter regions while maturation of neuroprogenitors in the hippocampus occurs. The presence of activation (e. g. H3K4me3 and H3ac) or repression (H3K9me3) epigenetic marks on histone H3 proteins associated with this promoter region is also shown.

On the other hand, we have continued analyzing the epigenetic mechanisms that control the expression of master regulators during osteoblast differentiation. Through a doctoral thesis co-directed by Drs. Montecino and Allende (Mrs. Adriana Rojas) and the work of a post-doctoral fellow (Dr. Gino Nardochi), we have established that different DNA sequences distributed along the Runx2 (bone master regulator) gene locus exhibit defined patterns of epigenetic histone modifications and enrichment of chromatin regulatory proteins. As transcription of Runx2 is stimulated in pluripotent mesenchymal cells by the morphogen BMP2, we also determined that this activation is accompanied by chromatin remodeling at the promoter region and changes in epigenetic modifications at histones H3 and H4 associated with this sequence. Moreover, we determined the contribution of lysine-methylase complexes containing WDR5, EZH2 and PMRT5 as well as lysine-demethylase

complexes including UTX, NO66, JARID1B and JARID1C during transcriptional regulation of the Runx2 gene in mesenchymal differentiation. Silencing of WDR5 and UTX in BMP2-stimulated cells decreases Runx2 expression. Also, WDR5 and UTX are bound to the Runx2 promoter in cells treated with BMP2. Together these results demonstrate that WDR5 and UTX are important components of regulation of Runx2 transcription during mesenchymal differentiation by establishing an epigenetic environment at the transcriptional start site that facilitates binding of the transcription machinery (Rojas et al. manuscript in preparation).

b) Development, stem cells and regeneration

Molecular genetics of sensory cell regeneration.

We use the zebrafish embryo and larva to study how neurons and sensory cells in a vertebrate can regenerate when damaged. We have reported on a novel technique that will allow us to generate tissue and cellular damage in any part of the sensory system using electroablation (Moya et al., 2014). Using this method, we have shown that a population of stem cells responsible for organ regeneration resides outside the organ and can migrate towards the damaged area to participate in repair. We have further shown a crucial role for glial cells (specifically, Schwann cells) in both axonal and sensory cell regeneration in this system. The next step will be to use next generation sequencing to determine which genes are turned on in both the stem cells and in the Schwann cells in response to damage and to perform a comparison with similar studies done in human tissues. This should provide clues as to how fish -with a similar repertoire of molecular players- can regenerate so well in comparison to mammals.

Development of animal models for high-throughput small molecule screens.

Our work in immunology and regeneration has been complemented by the development of assays that can be used for drug screens, efficacy of treatments for various disease states and toxicology using the zebrafish model. As in previous years, one of the most successful areas in which these technologies can be applied is in the area of cancer biology. Using an angiogenesis model in zebrafish larvae, we were able to prove *in vivo*, that coagulation factor X mediates an important antiangiogenic response, and that this activity is conserved through mammmals (Lange et al., 2014). Once again, we will pursue the use of this and other animal models as discovery tools for biomedicine and environmental studies.

Sonic Hedgehog signaling in the developing optic tectum of vertebrates: A comparative perspective

Precise coordination of pattern formation and proliferation is essential for normal brain morphogenesis. Apart from its early morphogenic role in the ventral regions of the CNS, the Shh/Gli pathway also has an important histogenetic function by regulating stem cell proliferation in the associative areas derived from the alar plates of the mammalian brain. This led us to investigate its possible involvement in the control of stem cell lineages in the dorsal midbrain (optic tectum) at late embryonic stages using a comparative approach. By taking advantage of developmental genetics combining a detailed analysis of the different zebra fish *Shh* and *Gli* mutants, with an analysis of Shh pathway gain of function in transgenic mouse models and *in vitro* neurosphere assays we characterized the contribution of Shh in OT development. In addition, we perturbed the hedgehog pathway in the midbrain *in ovo* by pharmacological gain and loss of function treatment and by Shh and Gli proteins electroporation in the developing chick OT. Our results show that the SHH/Gli pathway has an evolutionarily conserved role in controlling stem cell behavior in the developing vertebrate OT.

Importantly, our results provide comparative data to our current understanding of progenitor/stem cell mechanisms that place Shh as a key niche factor in the dorsal brain (Feijóo et al., *EJN* 2011; Rapaciolli et al, *BMC Neuroscience, 2012*; Martínez et al., *PLOS ONE 2013*).

The Neogenin 1 (Neo1) receptor mediates Sonic Hedgehog (Shh) driven neural precursor cell proliferation and tumor growth

The growth factors that regulate proliferation and maintenance of stem cells are responsible for maintaining homeostatic proliferative balance, that when broken, could trigger a malignancy. As such aberrant Shh pathway activation has been associated with various cancers such as medulloblastoma (MB) and neuroblastoma (NB) two common childhood malignant tumors of the cerebellum and sympathetic nervous system, respectively. By using different genomic approaches we have recently uncovered Neo1 as a direct Shh target (Milla et al., *BMC Genomics*, 2012). Neo1, a death dependence receptor classically known as a Netrin1/RGM binding partner, has been involved in many processes during nervous system development. We therefore have been investigating the involvement of Neo1 mediated Shh signaling in the pathogenesis of MB and NB. Our results show that Neo1 is regulated by the canonical Shh signaling in granule neuron precursor proliferation in the developing cerebellum and may play a prominent role in the onset of MB, an observation essential for improving pediatric anticancer pharmacology (Milla et al., *International Journal of Cancer 2013*). Ongoing research aims to extend these findings studying Neo1 overexpression as a biological marker for NB (Arros et al, manuscript in preparation).

Sonic Hedgehog modulates EGF responsiveness through EGFR transactivation in neural precursor cells.

Shh and EGFR signaling pathways modulate Neural Stem Cells (NSC) proliferation. How these signals cooperate is therefore critical for understanding normal brain development and function. During 2013 we reported a novel acute effect of Shh signaling on EGFR function. We showed that during late neocortex development Shh mediates ERK1/2 signaling pathway activation in Radial Glial cells through EGFR transactivation. These findings may have important implications for understanding the mechanisms that regulate NSCs proliferation during neurogenesis (Reinchisi et al, Frontiers in Cellular Neuroscience, 2013). Importantly, recent evidence has also involved Shh/Gli and EGFR cooperative interaction in oncogenic transformation and therefore our results may lead to novel approaches to the treatment of brain tumors, a matter currently under investigation.

Role of Neo 1 and Netrins as non-classic angiogenic molecules in Tumor Vessel Vascularization and its modulation by Shh signaling

Unwanted neovascularization is known to contribute to tumor progression and metastasis. Despite progress in understanding the molecular basis of angiogenesis, and successful VEGF blockade for the treatment of some cancer patients, challenges must be overcome to improve the overall efficacy of antivascular strategies to combat cancer more efficiently. Netrin proteins and their guidance receptors, Neo1 and DCC (Deleted in Colorectal Cancer), regulate cell and axon migration. But also they have recently implicated in tissue morphogenesis, tumorigenesis and angiogenesis. However, the role of Netrins/Neo1 has not been explored in detail and so far has not been related at all to Shh signaling. The aim of this work is to elucidate the role of Netrins in the induction of angiogenesis acting through their putative receptors, focusing in Neo1, with special emphasis in the formation/ maintenance of Shh derived cancers.

In collaboration with Dr Allende and Dr Owen we have been studying the rol of coagulation factor Xa in angiogenesis in vitro and *in vivo* (Lange et al., J Cell Physiol. 2014). Recently these results have been extending to analysis of Fxa contribution in tumour growth and metastasis presenting the first demonstration for FXa in cancer progression.

Mesenchymal stem cells (MSCs) as therapeutic agents for both tissue engineering and immunotherapy applications
Disorders in skin wound healing are a major health problem that requires the development of innovative treatments. The use of biomaterials as an alternative of skin replacement has become relevant, but its use is still limited due to poor vascularization inside the scaffolds, resulting in insufficient oxygen and growth factors at the wound site. Our ongoing studies are related to the angiogenic and immunoregulatory effects of human umbilical chord MSCs. In particular, in collaboration with Dr Egaña, we have developed a cell-based wound therapy consisting of the application of collagen based dermal scaffolds containing mesenchymal stem cells from Wharton's jelly (WJ-MSC) in an inmmunocompetent mouse model for dermal regeneration (Edwards et al., submitted to *Angiogenesis*). Although it is clear that the therapeutic approach that we propose seems promising, current research is carried out in order to evaluate the molecular mechanisms that regulate the various events of wound healing and the particular contribution of WJ-MSC to vascular function. In particular we have studied how Shh improves WJ-MSC mediated angiogenesis *in vitro* e *in vivo* focusing on the effect over VEGF and Ang-1 (Angiopoietin-1) expression levels. Our results demonstrate that Shh signaling acts on WJ-MSC stimulating *VEGF* and *Ang*-1 expression, resulting in a beneficial effect on the angiogenic activity of WJ-MSC, which might be helpful in accelerating the angiogenic process (Zavala G. & Palma V, manuscript in preparation).

Role of the cystein-serine rich nuclear proteins (CSRNP) in progenitor cell expansion and survival.
This is a collaborative effort between Dr. Glavic and Dr. Allende. Our work has shown that, both in Drosophila and in Zebrafish, members of this family (DAxud1 and csrnp1 respectively) regulate the proliferation and survival of precursors at different tissues (Glavic et al., 2009; Feijoo et al., 2009; Espina et al., 2013). We recently obtained the RNA-seq data from loss and gain of function condition of DAxud1 and we will use it, together with the result from in vivo DamID experiments in D. melanogaster (Southall et al., 2013; Dev Cell 26:101-112) to identify its target genes and to gain understanding about the conserved mechanisms used by these proteins to perform their proliferative functions.

c) The stress response: genomic and proteomic outcomes
Biotic and Abiotic stresses are related to the activation of transcription factors associated to stress in the endoplasmic reticulum.
bZIP60 is the functional homologue to HAC1 and XBP1 in yeast and animals respectively. These genes are transcription factors that respond to ER stress through the IRE1 signaling pathway, which is activated during the unfolded protein response. The activation of IRE1 leads to the unconventional splicing of the mRNA of these transcription factors. Interestingly, whereas the unspliced form of HAC1 and XBP1 seem to be not functional, our results indicate that the unspliced form of bZIP60 in Arabidopsis thaliana is related to the plant response to abiotic stress. We found that plants grown in mannitol to promote osmotic stress show no splicing of bZIP60, but we observed translocation of this transcription factor fused to GFP to the nucleus of plant cells under these conditions. We are currently

looking for the target genes of the unspliced form of bZIP60 using transcriptome analyses of wild type and mutants in bZIP60 grown in the presence and absence of mannitol. In addition, we are doing chromatin immunoprecipitation analyses in order to identify target genes.

Molecular mechanisms operating during cellular stress in Drosophila cells.
The unfolded protein response (UPR) is a stress response evoked by the accumulation of misfolded proteins at the ER. The KEOPS/EKC complex has been associated with a tRNA modification that ensures correct codon recognition and proper translation. We have performed in vivo disruption of this complex showing that indeed it induces the UPR affecting also TOR activation and cell and tissue growth (Ibar et al., 2013; Rojas-Benítez et al., 2013). Additionally we have started a collaboration with Dr. González to characterize the role of PI3K/Foxo pathway in the homeostatic behavior of *Drosophila* fatbody cells (the analogue to adipose and hepatic tissues from mammals) exposed to different amounts of copper and other metals.

d) Networks and modeling

Systems analysis of transcriptome data provides new hypotheses about *Arabidopsis* root response to nitrate treatments.

Nitrogen (N) is an essential macronutrient for plant growth and development. Plants adapt to changes in N availability partly by changes in global gene expression. We integrated publicly available root microarray data under contrasting nitrate conditions to identify new genes and functions important for adaptive nitrate responses in *Arabidopsis thaliana* roots. Overall, more than two thousand genes exhibited changes in expression in response to nitrate treatments in *Arabidopsis thaliana* root organs. Surprisingly, 60% of differentially expressed genes were regulated by nitrate in only one experiment, indicating nitrate regulation of gene expression depends largely on the experimental context. However, analysis of regulated biological functions showed that responses at the functional level are more robust from experiment to experiment as compared to genes. The average number of genes shared between any two experiments is 6.7%, while 19.5% of overrepresented GO terms (FDR <0.05) are shared in the same two experiments. This difference in the percentage of shared genes versus GO terms increases with the number of experiments compared. For example, the number of GO terms shared between any five experiments is ten times higher than the number of shared genes. This difference between intersection of genes and GO terms was not due to an artifact due to the nature of the gene to GO association data. First, conservation of GO terms could not be explained by genes annotated to very general GO term categories increasing the chance of intersection at the GO term level. The distribution of levels is similar for total and shared GO terms indicating differentially expressed genes are not biased towards general GO term categories. Second, the average number of GO terms associated to nitrate-responsive genes is very similar to a random sample of genes taken from the Arabidopsis genome. This result rules out the possibility that nitrate responsive genes have more annotations than a random set of genes of the same size therefore increasing the chance of an overlap. Third, GO terms shared between any two experiments contained on average only 22.4% of the same genes. This result indicates most of the genes contributing to over-represented GO terms are different in each experiment. Finally, a prediction of our hypothesis is that different members of the same gene family should be found contributing to shared GO terms between different experiments. To systematically test this hypothesis, we performed pair-wise comparisons of protein sequences for all

genes annotated to shared GO terms between any two experiments. We then compared the distribution of protein sequence similarities for all pair-wise comparisons between genes contributing to a shared GO term in our data set versus all genes annotated to that GO term. We found shared GO terms (74.5%) have more pairs of similar protein sequences coming from different experiments than expected by chance ($\alpha < 0.05$). For example, *NRT2.2* gene was found differentially expressed in experiment 21 and *NRT2.5* was found differentially expressed in experiment 23. Both *NRT2.2* and *NRT2.5* are annotated to the shared GO term "Transport (GO:0006810)". Similarly, the shared GO term "Response to Carbohydrate (GO:0009743)" contains *GLN1;2* and *GLN1;1*, each regulated in different experiments. The easiest interpretation of these results is that nitrate responses at the biological function level are more robust to experimental context than genes. This phenomenon could be explained by functional redundancy of different genetic components, a feature that is common to biological networks and has been proposed as a mechanism for robustness toward stochastic fluctuations (Whitacre, 2012). A similar idea is the degeneracy concept proposed by Edelman and Gally (2001), which defines the property whereby structurally different elements may perform the same or similar functions. This feature has been attributed not only to gene networks but also to neural networks and evolution (Edelman and Gally, 2001). This phenomenon may be particularly relevant in plants, where increased gene family sizes may provide higher adaptive capacity to environmental perturbations.

Integrative gene network analysis uncovered relationships between nitrate-responsive genes and eleven highly co-expressed gene clusters (modules). Four of these gene network modules have robust nitrate responsive functions such as transport, signaling and metabolism. To identify transcription factors that control essential and robust functions in the root nitrate response such as nitrate transport and assimilation, we focused in transcription factors from these modules and their possible targets (Figure XXX). In this network, MYB-related (AT5G58900) and bZIP (AT5G65210) genes showed the highest degree. Three different G2-like transcription factors (AT1G68670, AT1G25550, AT1G13300) were also found in top positions of the ranking of transcription factors with higher degree. MYB-related gene coexpressed with nitrite reductase, 2-oxoglutarate/malate chloroplast transporter and a 6-phosphogluconate dehydrogenase gene from the oxidative pentose phosphate pathway. These results suggest this MYB-related factor controls basic aspects of nitrate metabolism, such as nitrate reduction, GS/GOGAT cycle and the generation of reducing equivalents. A bZIP transcription factor identified as potential network driver (*TGA1*) belong to the subfamily TGA and another member of this family (*TGA4*) occupied the fifth position in the ranking of transcription factors with higher degree. These transcription factors have been implicated in bacterial defense responses. *tga1/tga4* double mutant plants show a greater susceptibility to infection by *Pseudomonas syringae* (Shearer et al., 2012). However, our analysis suggests that these transcription factors (*TGA1*, *TGA4*) could be important in the nitrate response of *Arabidopsis* roots.

Integrated network analysis of transcriptome data provided novel hypothesis about functions and regulatory mechanisms by which *Arabidopsis* plants respond to nitrate. Our meta-analysis better assessed the nitrate functional space than any single or integrated transcriptome study previously published. We estimated the mean functional coverage of any single experiment at about 31%. This result highlights the need for integrated data analysis to better map the functional space for any given perturbation. Moreover it underscores the need for using experiments carried out under non-redundant environmental conditions.

Our Systems approach identified nitrate regulation of root hairs as an important component of the plant developmental response to changes in N nutrition, a yet unexplored research area at the

intersection of N nutrition and root biology. We provided concrete hypothesis for genes and connections among genes related to root hair differentiation in response to nitrate that have not been previously highlighted nor addressed experimentally. Our results also highlight the role of bZIP and G2-like transcription factors for regulation of important functions related to nitrate transport and signaling. G2-like transcription factors have not been characterized in the context of nitrate responses. Functional studies of these new candidate genes should help better understand regulatory mechanisms underlying root nitrate responses in *Arabidopsis* and other plants.

Reconstruction of regulatory networks

Abundant evidence shows that abiotic stresses strongly affects gene expression, yielding different forms of gene synchronization. A natural explanation is the existence of shared transcriptional regulators. However, the discovery or complete characterization thereof remains a challenging problem. State-of-the-art bioinformatics regulation discovery methods are commonly based on the prediction of putative regulator genes and their targets, together with different ways of integrating predictions of co-regulated genes from expression data. Nevertheless, in general, this integration generates lists of results which are orders of magnitude greater than experimentally validated data and needs thorough post processing. During 2013 we proposed a method that integrates predictions of transcription factors, binding sites and operons with gene associations induced by transcriptomic data in order to produce a realistic regulatory graph. This graph results from the solution of a combinatorial problem implemented using Answer Set Programming, a logic-based paradigm which enables an effective encoding and solving of complex combinatorial problems. Our approach, benchmarked on *E. coli*, provided a regulatory graph that recovers essential features of the gold standard regulatory network for this organism, keeping experimentally validated regulations with significantly higher probability than non-validated ones. In addition, it shares the expected topological properties of a regulatory network. As a major functional output, our approach can be used to highlight functional relationships between genes clustered together in transcriptomic experiments but moreover emphasizes within the whole genome the key functional global regulators which are necessary when the bacterial system is stressed by environmental conditions. This work has been submitted to Plos Computational Biology.

In addition to the afore mentioned reconstruction method we study the computational complexity of such problem. When the problem is modeled as the minimization of a global weight function, we show that the enumeration of scenarios is a hard problem. As an heuristic, we model the problem as a set of independent minimization problems, each solvable in polynomial time, which can be combined to explore a relevant subset of the solution space. We present a logic- programming formalization of the model implemented using Answer Set Programming. We show that, when the graph follows patterns that can be found in real organisms, our heuristic finds solutions that are good approximations to the full model. This work is accepted for oral presentation in the 15th International Conference on Verification, Model Checking and Abstract Interpretation and will be publisehd in the Lecture Notes in Computer Sciences.

Reconstruction of metabolic networks

Genome-scale metabolic models are a powerful tool to study the inner workings of biological systems and to guide applications. The advent of cheap sequencing has brought the opportunity to create metabolic maps of biotechnologically interesting organisms, that are often related to a studied

reference organism. While this drives the development of new methods and automatic tools, network reconstruction nonetheless remains a time-consuming process and extensive manual curation is required. This curation introduces specific knowledge about the modeled organism, either explicitly in the form of molecular processes, or indirectly in the form of annotations of the model elements. Paradoxically, this knowledge is usually lost when reconstruction of a different organism is started. By combining a knowledge base encoded in an annotated SBML scaffold model, orthology mapping between genes, and experimental phenotypic evidence, we built a genome-scale metabolic model of the target organism that is well suited for manual curation. In particular our method infers implicit knowledge from the annotations in the scaffold, and rewrites these inferences to include them in the resulting model of the target organism. Scripts for evaluating the model with respect to experimental data are automatically generated, to aid curators in iteratively improvement.
As an application, using the transcriptome of *Nannochloropsis salina* iNR890 we have produced the first genome-scale metabolic model for this microalga. It includes 1,927 reactions, related to 890 genes and distributed into eight compartments. We identified metabolic pathways and reactions that are critical for this species survival. We based our reconstruction on an existing model of the alga *Chlamydomonas reinhardtii*. Our results show both common and exclusive metabolic reactions between both algae. This is the first well-annotated model of a microalga, providing a reference for future metabolic improvements, a guide to future experiments and a starting point for the metabolic reconstruction of other algae. In recent times, *N. salina* has gained great scientific and biotechnological relevance due to their potential in generating biofuel and supplements like Omega 3 for the food industry.

Dynamics of metabolic networks

The increasing availability of metabolomics data enables to better understand the metabolic processes involved in the immediate response of an organism to environmental changes and stress. The data usually come in the form of a list of metabolites whose concentrations significantly changed under some conditions, and are thus not easy to interpret without being able to precisely visualize how such metabolites are interconnected. We developed a method that enables to organize the data from any metabolomics experiment into metabolic stories. Each story corresponds to a possible scenario explaining the flow of matter between the metabolites of interest. These scenarios may then be ranked in different ways depending on which interpretation one wishes to emphasize for the causal link between two affected metabolites: enzyme activation, enzyme inhibition or domino effect on the concentration changes of substrates and products. Equally probable stories under any selected ranking scheme can be further grouped into a single anthology that summarizes, in a unique subnetwork, all equivalently plausible alternative stories. An anthology is simply a union of such stories. We detail an application of the method to the response of yeast to cadmium exposure. We use this system as a proof of concept for our method, and we show that we are able to find a story that reproduces very well the current knowledge about the yeast response to cadmium. We further show that this response is mostly based on enzyme activation. We also provide a framework for exploring the alternative pathways or side effects this local response is expected to have in the rest of the network. We discuss several interpretations for the changes we see, and we suggest hypotheses that could in principle be experimentally tested. Noticeably, our method requires simple input data and could be used in a wide variety of applications. This work was accepted in the journal Bioinformatics.

Proposing functional metabolic unities

Integrating heterogeneous knowledge is necessary to elucidate the regulations in biological systems. In particular, such an integration is widely used to identify functional units, that are sets of genes that can be triggered by the same external stimuli, as biological stresses, and that are linked to similar responses of the system. Although several models and algorithms shown great success for detecting functional units on well-known biological species, they fail in identifying them when applied to more exotic species, such as extremophiles, that are by nature unrefined. Indeed, approved methods on unrefined models suffer from an explosion in the number of solutions for functional units, that are merely combinatorial variations of the same set of genes. We have overcome this crucial limitation by introducing the concept of "genome segments". As a natural extension of recent studies, we rely on the declarative modeling power of answer set programming (ASP) to encode the identification of shortest genome segments (SGS). This study shows, via experimental evidences, that SGS is a new model of functional units with a predictive power that is comparable to existing methods. We also demonstrate that, contrary to existing methods, SGS are stable in (i) computational time and (ii) ability to predict functional units when one deteriorates the biological knowledge, which simulates cases that occur for exotic species. This work was accepted for oral presentation in the 12th International Conference on Logic Programming and Nonmonotonic Reasoning and was published in the Lecture Notes in Computer Sciences.

Biological acclimatization using hybrid systems

In order to describe the dynamic behavior of a complex biological system, it is useful to combine models integrating processes at different levels and with temporal dependencies. Such combinations are necessary for modeling acclimatization, a phenomenon where changes in environmental conditions can induce drastic changes in the behavior of a biological system. A modeling scheme called *strong switches* is proposed. It formalizes the use of hybrid systems as a tool to model this kind of biological behavior. We have illustrated the proposed methodology with two applications: acclimatization in wine fermentation kinetics, and acclimatization of osteo-adipo differentiation system linking stimulus signals to bone mass.

Further in this line, we propose a hybrid model describing the multi-cell dynamics at different levels in yeast. The model integrates regulatory mechanisms, which are associated to gene regulation with those associated to epigenetics. Importantly, our model includes the epigenetic inheritance effect due to changes in calorie restrictions, the effects of aging and spatial constraints, and we give insights about the influence of each model coefficient on the stimulus signals of proliferation. The model is highly stiff and exhibits frequent discontinuities, which makes it very expensive to simulate using conventional numerical algorithms. To overcome this problem, we implemented it in a solver based on quantized state system methods, LIQSS2, which efficiently handles discontinuities and stiffness. This is the first model describing the precursor cells in yeast, taking into account several regulatory mechanisms and epigenetic inheritance, giving origin to asymmetric cell divisions in relation to the epigenetic mark H3K9Ac. The model allows us to predict the dynamic of yeast cells and the epigenetic mark H3K9Ac, and explore ways to stimulate the proliferation of epigenetically marked cells.

Bioinformatic survey for new physiological substrates of Cyclin-dependent kinase 5

Cyclin-dependent kinase 5 (Cdk5) is a proline-directed serine/threonine kinase predominantly active in the nervous system where it regulates several processes such as neuronal migration, cytoskeletal dynamics, axonal guidance, and neurotransmission. We constructed a position specific scoring matrix (PSSM) based on a dataset of sites shown to be phosphorylated both in vivo and in vitro by Cdk5.

This dataset was curated manually through an exhaustive search of published experimental data. We then used this PSSM to perform a search in the mouse proteome through Scansite, a web-based tool for matching sequence patterns in large databases. Considering a stringent cut-off score of 0.5, we identified 354 new putative sites present in 291 proteins. In order to assess the robustness of our results, ten random subsets (of 80 sites each) of the original dataset were used to construct new PSSMs, which were then used as input for a new Scansite search, leading to the recovery of 81% of the 354 sites by at least 5 PSSMs.
In order to reduce the number of false positives in our sequence-based approach, we evaluated which of these predicted sites were phosphorylated in vivo as determined by multiple phosphoproteomics studies carried out through mass spectrometry and available in the PhosphoSitePlus database. This step resulted in a very promising list of 132 putative phosphorylation sites for Cdk5, of which, 51 are specifically phosphorylated in brain tissue, and some are involved in functions regulated by Cdk5 such as axonal growth, synaptic plasticity and neurotransmission. Other phosphorylation sites in our list suggest that Cdk5 might regulate processes through mechanisms not previously recognized such as the control of mRNA splicing.

ii. Synergy and collaboration among research lines

As was stated in the previous report, synergy between the investigators and specialties is achieved by defining Center or core projects that receive scientific and intellectual input from multiple (often all) Principal and Associate Investigators. As we have decided that these objectives must, by definition, be collective efforts, and because they receive speicifc funding, investigators and members of their groups become involved in all stages of their execution. Thus, we expect a growing number of co-authored publications, something that was non-existent before the grant and has still not manifested itself to the extent we desire. Core or Center projects are discussed in at least three formats: first, we hold weekly meetings of all the Pis to closely follow the progress of each objective or project. Second, we have monthly seminars called "Interactomics", in which we alternated between invited speakers and work in progress tech talks that focus specifically on the core projects. Thirdly, we have had another CGR retreat this year in which all PIs and AIs discuss the state of the projects and new strategies to approach them. A fourth avenue for generating cross-disciplinary projects will be implemented this year (a suggestion made by our reviewers) which will consist in retreats where only postdocs and graduate students will have the chance to present their research to each other. It is often the case that young investigators are more prone to seek out new research projects given their direct involvement with experimental science.
The Center projects are maturing to the stage of a first wave of publications, where multiple authorship of PIs will become more frequent. We have increased the number of papers co-authored by two or more CGR investigators to 8 in 2013, compared to 3 in 2011 and 4 in 2012. This is a trend we expect to sustain towards the final stage of the first five years of funding and is a reflection of our commitment to making the CGR a truly collaborative and focused center.

iii. Formation of advanced human capital directly related to the Center´s objectives

We continue to privilege the training of advanced human capital at all levels of academic and professional development. In addition to our traditionally strong numbers of trainees, we have also

invested effort in the organization of courses, seminars, workshops and tutorials aimed specifically at CGR and external young scientists. Thus, the impact of our center on the new generation of genome scientists has increased substantially. As has been the case in the previous years, students are continually attracted to work in the center labs and the high quality of these recruits can be demonstrated with the high success rate of their applications to fellowships. In fact, very few of our students and postdocs receive stipends directly from the CGR, as they are very likely to obtain outside funding. As an example, of the six CGR postdoctoral fellows that applied for FONDECYT grants in 2013, all six succeeded.

Also of note is that one of our former postdocs, Dr. Christian Hödar, was incorporated as an Associate Investigator in the CGR. We are interested in carreer development of all of our scientists and we have aimed at making sure they can enter the job market with strong qualifications.

One aspect we hope to strengthen this upcoming year is the identification of the students and postdocs with the central objectives and spirit of the CGR. The international evaluation panel recommended that we organize activities oriented at young investigators, where they could take charge and lead with proposals for stimulating discussions and interactions. We have endowed both graduate students and postdocs with this mandate (and funds) and they are in the organization stages of activities such as retreats, invitation of foreign speakers and social gatherings. We hope this will stimulate even further the synergy we seek among our labs.

iv. Collaborative networks both at the national and international level

This past year we established several important contacts with international institutions. We organized two workshops, one with the University of Tokyo and another with the University of Heidelberg. In both cases, foreign scientists came to Chile to participate jointly with CGR and other national researchers in scientific sessions that will lead to joint proposals for student and academic exchange. We also initiated collaborative work with investigators from Harvard University and the Pasteur Institute in Montevideo, Uruguay. Two more memorandums of understanding that will lead to full fledged agreements are under review, one with the Universidad Pablo de Olavide in Sevilla, Spain, and one with Monash University in Melbourne, Australia. As in previous reports, many of our papers are published together with foreign collaborators (15 articles in this report).

Many of our collaborations in genomics have been established with local colleagues, such as Dr. Claudio Latorre, a specialist on the flora of the northern *Altiplano*, Dr. Marco Mendez, on the biology of the *Orestias* fish and the *Rhinella* amphibians, Drs. Reinaldo Campos and Fransica Blanco, on the flowering desert, and Dr. Patricio Hinrichsen, on the Sultanina grape. The Chilean Human Genome Project has also involved interaction with other Chilean investigators and we have recently proposed to establish a national Consortium to handle the large amount of data generated and to make it available to the public, especially to Physicians and geneticists that might use it. We plan to use the CGR platform in Bioinformatics (with a significant hardware upgrade) to host the database that will arise from this pooling of data.

Our networking efforts respond to the strategic objective stated in the original proposal aimed at strengthening the capacity of our nation to embark in genome projects in any species of interest. Without the CGR, many these projects would not exist today.

v. Dissemination and exploitation of results

Dissemination was carried out principally in publications, congresses and conferences given by CGR researchers. The details of these instances can be found in the attached tables.

vi. Outreach to society

We provide an extensive overview of our activities towards the public in Appendix B. In summary, our efforts are aimed at exposing the public in diverse media to genome science and to the different topics that we work on at the CGR.

b. Describe unexpected difficulties encountered and indicate how they were dealt with.

We have not experienced significant difficulties at the scientific or administrative levels. We are aware that many of the Center projects are taking longer to mature than anticipated. The main reason for this (a problem that is shared with simialr centers around the world, it seems) is the bottleneck at the level of data analysis. Since we are dealing with novel or complex species when we analyze genomes or transcriptomes, we need very well trained bioinformaticians to tackle the projects. Right now, there are very few people qualified in Chile to do this work, almost all of them trained by Dr. Maass. Despite the fact that there has been a strong recruitment of new personell in this area (engineers and biotechnologists), there is still a lack of informaticians well versed in the language of DNA/RNA. To ameliorate this problem, we have embarked on an aggressive campaign to recruit and train new people in this area. To start, we have set aside funds for hiring three new full time bioinformaticians that will deal exclusively with CGR core projects. They should be able to satisfy the current demands for manpower and accelerate some of the projects that are not advancing as fast as they could. Nonetheless, the interaction of biologists and engineers/mathematicians is still required for the correct application of scripts to the data, so it is also inherent to the nature of our research that some results will take time to see publication. We do think that the main results of all of our research lines will be pubished by the five-year funding period, being 2014 and 2015 the most productive years for the CGR within this timeframe.

2. RESULTS ACHIEVED PER RESEARCH LINE

Briefly describe the main results per research line achieved during the period.

Since we have described the results obtained thus far in each of the Research Lines above, in this section we will simply summarize the main publications related to each of the Lines. We also follow up with a current total of productivity metrics for the project in order to follow the progress of the Center towards becoming highly relevant in the field.
The main advances per Research Line are:
Line 1. Multiple altiplano plant transcriptomes obtained (article to be submitted at the end of 2014); transcriptome of Rhinella amphibians in diverse ecological contexts complete and article submitted early 2014; genome of first Orestias fish complete, five more species in progress for an article to be submitted in 2015; Transcriptome of annual flowering desert plants complete; Transcriptome of annual fish at diverse embryological stages complete and under analysis; metagenomics of desert microorganisms under way.
Line 2. Article describing the Mapuche/Huilliche genome has been submitted; we include this article as Appendix A. Other genomes of relevance that were completed are: the *Sultanina* table grape genome published (two articles), the Atlantic salmon (annotation of the genome in progress; article expected at the end of 2014) and the pathogen *Pitsirickettsia*, where we have completed sequencing and comparative analysis.
Line 3. Most of the experimental work, and thus the publications generated by the CGR, fall within this line as it encomapsses the more traditional areas of research we follow. The most relevant publications were generated in the areas of stem cell research, cancer biology, regeneration, tissue engineering, network modelling, mathematical theory and basic cellular and molecular biology.

Indicator	2011 (yr1)	2012 (yr 2)	2013 (yr 3)	Accumulated or average
Number of ISI papers	34	35	37	106
Total Impact Factor of ISI papers	177.9	161.4	184.1	523.4
Average Impact factor of ISI papers	5.2	4.7	5.0	4.96
5 year citations (papers); average per paper*	1608 (153)	1716 (153)	1657 (129)	10.54
Co-authored publications#	3	4	8	15
Postdocs associated to CGR&	24	26	32	27
PhD students associated to CGR&	43	42	45	43
Total number of theses directed&	84	82	93	86
Co-directed theses& (CGR PIs)	5	4	7	5
Congress presentations	140	118	155	413
Conferences and courses organized	9	9	12	30

*Citations for articles published by the 6 Principal Investigators with a window encompassing the five previous years
#Papers in which more than one CGR investigator (Principal or Associate) are authors.
& As many students are the same from year to year, the numbers appearing in the "Accumulated" column are averages rather than sums

IV. SUGGESTIONS FROM PREVIOUS EVALUATION

Describe how the suggestions provided by the evaluation panel and the FONDECYT Council in its previous evaluation report were taken into account by the Center.

In June 2013 we had the site visit of a new evaluation committee composed of Dr. Jörg Hoheisel of the German Cancer Research Center and Dr. Igor Stagljar, of the University of Toronto. They met for an entire day with the PIs, the AIs, the postdocs and students of the CGR. They generated a report to which we had access in July, as well as a summary letter from the FONDAP program. Overall, the evaluation is very positive (all grades were "very good" or "outstanding") and expresses optimism regarding the Center's future. The reviewer's report indicates that progress and productivity are well in excess of what can be expected given the funding provided. They also note that the choice of center projects has been correct, as focusing on objectives that have a "Chilean flavor" ensures that the CGR will have a place in world science.
The panel made some recommendations. First, they made clear that objective 3 of our project should remain as one of the Center's priorities (a core objective), rather than becoming subservient to objectives 1 and 2. We understand their opinion and agree that it is the strongest and most productive area of the science done at the CGR, despite the interest that objectives 1 and 2 can generate locally. We will continue to provide this area the status it deserves in our Center and it will show increasing integration to the other research lines. Second, they made two recommendations regarding young scientists at the CGR. Postdocs and PhD students should have an instance where they can gather independently of the Main Investigators and carry out scientific exchanges "at the grassroots level". We have agreed this is an important oversight on our behalf and we have separated funds in 2014 and we have asked specific individuals to organize such events. Finally, they suggested the students should have a role in organizing the visit of speakers of their choice from abroad. We have also allocated funds for this purpose in 2014 and have designated a group of representative students to carry out this initiative.

V. PRODUCTS GENERATED BY THE PROJECT

In what follows, complete the attached Excel spreadsheets taking into account the following:

REPORT ONLY PUBLISHED MATERIAL INCLUDING THOSE WITH AN OFFICIAL DOI POINTER (e.g., with EARLY ONLINE ACCESS).

EXCEPT FOR BOOKS, ALL BACKUP DOCUMENTS MUST BE PRESENTED IN DIGITAL FORMAT. DO NOT SEND PRINTED COPIES.

ONLY PUBLICATIONS THAT ACKNOWLEDGE THE FONDAP PROGRAM WILL BE CONSIDERED.

1. ISI Publications

- ✓ For each publication, if applicable, the principal author and the corresponding author must be indicated using the following terminology:
  - [1] For principal author (example: Toro[1], J.)
  - [2] For the corresponding author (example: Toro[2], J.)
  - [3] For principal and corresponding author (example: Toro[3], J.)
- ✓ Include a digital copy of each **PUBLISHED** paper.

2. Non ISI Publications

- ✓ For each publication, if applicable, the principal author and the corresponding author must be indicated using the following terminology:
  - [1] For principal author (example: Toro[1], J.)
  - [2] For the corresponding author (example: Toro[2], J.)
  - [3] For principal and corresponding author (example: Toro[3], J.)
- ✓ Include a digital copy of each **PUBLISHED** paper.

3. Books and book chapters

- ✓ Include a hard copy of every **PUBLISHED** book.
- ✓ Include a digital copy of the front page of the chapter in the case of a book chapter.

4. Patents

- ✓ Include all patents generated by the FONDAP Center.

5. Congress presentations

- ✓ Include abstracts of all presentations. Attach a digital copy of the front page of the congress/workshop book.

6. Organization of Scientific Meetings

- ✓ List all congresses, courses, conferences, symposia, or workshops organized by the FONDAP Center.
- ✓ Include abstracts of all presentations. Attach a digital copy of the front page of the congress/workshop book.

7. Collaborative Activities

- ✓ List the scientific visits of Center members to international institutions
- ✓ List the scientific visits of foreign researchers to the Center in Chile.

8. Postdoctoral Fellows

- ✓ List postdoctoral fellows working in the Center during the reported period regardless of their funding sources.
- ✓ Provide current affiliation and positions held by former postdoctoral fellows that left the Center during the reported period

9. Students

- ✓ List titles of theses framed in the project completed during the reported period. Attach an abstract and the subject index.
- ✓ List titles of theses in progress, framed in the project, during the reported period. Include digital copies of the corresponding thesis registrations.
- ✓ Provide current affiliation and positions held by former students that graduated during the reported period

10. Funding Sources

- ✓ List all funding sources including FONDAP.

VI. OTHER ACCOMPLISHMENTS

Report articles or notes published in the media (provide URL links, if available), awards, prizes, etc.

We would like to inform that during 2013 several very important pieces of equipment were added to the existing infrastructure at the CGR. For instance, two new sequencers were purchased (Illumina MySeq and Illumina Hi SCAN). With funds from the University of Chile, a new animal facility was built at the Faculty of Science; this will be officially inaugurated in 2014. Finally, a new advanced microscope was obtained through adonation of the University of Heidelberg to Dr. Allende: it is a Digital Light Sheet Microscope (DSLM) that allows three dimensional reconstruction of organs or embryos (see the newspaper article in the Outreach section, Appendix B).

VII. SUGGESTIONS

What recommendations would you make to the FONDAP Program Office to improve the performance of the Center and the review process? Please describe.

APPENDIX

A. Manuscript by Vidal et al., submitted to Nature in January of 2014 (accompanying letter indicates that it has been sent out for peer review).

B. In the following pages, we include a description of the activities carried out in the area of outreach and dissemination of the Center's activities. This area is the responsibility of the CGR's journalist, Ms. Karen Meyer.