Comisión Nacional de Investigación Científica y Tecnológica – CONICYT

**FONDAP Centers of excellence in research Program**

**Final Five Year Report**
**2011-2015**

# Center for Genome Regulation (CGR)


Comisión Nacional de Investigación
Científica y Tecnológica – CONICYT

**FONDAP CENTERS OF RESEARCH PROGRAM**

**FINAL REPORT**

**FIRST FIVE-YEAR PERIOD**

# FONDAP CENTER FOR GENOME REGULATION (CGR)

Guidelines:

The report should be written following the format specified hereafter. Both a printed (report and excel spreadsheets) and an electronic version must be sent to the following address:

**PROGRAMA CENTROS DE EXCELENCIA FONDAP**
CONICYT
Moneda 1375, Floor 9
Santiago

E-mail: mcamelio@conicyt.cl
Phone: (56 – 2) 2435 43 27

For future inquiries, please contact:

**María Eugenia Camelio**
FONDAP Program Interim Director
E-mail: mcamelio@conicyt.cl


Comisión Nacional de Investigación Científica y Tecnológica – CONICYT

## I. PRESENTATION

**PERIOD COVERED:** From: January 2011 To: June 2015

| **NAME OF THE CENTER**<br><br>**FONDAP Center for Genome Regulation** | | **CODE**<br><br>15 09 00 07 |
|---|---|---|
| **DIRECTOR OF THE CENTER**<br><br>Dr. Miguel L Allende | **E-MAIL**<br><br>allende@uchile.cl | **SIGNATURE** |
| **DEPUTY DIRECTOR**<br><br>Dr. Martín Montecino | **E-MAIL**<br><br>mmontecino@unab.cl | **SIGNATURE** |
| **SPONSORING INSTITUTION**<br><br>Universidad de Chile | | |
| **SPONSORING INSTITUTION REPRESENTATIVE**<br><br>Prof. Víctor Cifuentes (Dean) | **E-MAIL**<br><br>**vcifuentes@uchile.cl** | **SIGNATURE** |
| **ASSOCIATED INSTITUTION(S)** (if applicable)<br><br>Pontificia Universidad Católica de Chile, Universidad Andrés Bello | | |
| **CENTER WEBSITE ADDRESS**<br><br>www.genomacrg.cl | | |

**DATE**: 10/7/15

## II. EXECUTIVE SUMMARY

Five years ago, the FONDAP Center for Genome Regulation (CGR) set for itself a list of strategic and scientific objectives that would significantly change the landscape of Chilean genomic science and biological research. At the midpoint of the projected 10-year period in which these goals were to be accomplished, we can say that we are in the presence of a completely new scenario. The capacity for undertaking complex genome projects in the country has indeed advanced to the point of being on par with that of other nations and centers of similar scope. Training of investigators, mostly at the doctoral and postdoctoral levels, has generated sufficient expertise in the CGR to meet the challenges that we have faced successfully. The technological capacity required for the projects have accompanied our development, both in terms of sequencing capacity and computing power. We can declare with confidence, that we are now **the** reference center for genome projects in Chile, fulfilling our main strategic goal. Scientifically, we were able to identify relevant biological problems that satisfied two important criteria: first, they address unanswered questions in fundamental biology with ramifications in evolution, genome architecture, gene regulation and epigenetics. Second, they involve organisms that are important to the country because of the significance of their biological heritage, the uniqueness of their environments and the special natural history they have led as species. We have complemented these studies with model or economically valuable organisms and with a landmark project on human genomics. All of the projects have advanced to the point of revealing interesting and novel features that will be of high impact, as is described below. Importantly, these Center projects were addressed with the input and active participation of all of the six Principal Investigators with the support of the Associate Investigators in specific instances. In this sense, it is noteworthy that CGR members have committed to give priority and to dedicate their main efforts towards the Center projects, a true measure of the impact of collective -as opposed to individual- funding. In other words, the funding for the CGR has generated new areas of research, inaccessible to lone researchers, and has spawned synergistic interactions between investigators of different backgrounds, justifying their grouping under one scientific umbrella institution. The impact of the CGR has been felt in other, perhaps more expected, ways. As has been predicted, training has been one of our strongest indicators. The number of postdocs, students and professionals that have worked at the CGR is in the hundreds, which attests to the attractiveness of the Center to young scientists from all over the world. This has been complemented with courses, workshops, conferences and our internal dissemination of the science carried out at the CGR, of which these young people have been the main beneficiaries. We have not forgotten our responsibility towards the nation and society and we have made substantial efforts to promote our science and to educate on the subjects of our expertise. Our investigators are all leaders in their own right and many of us have been involved in policy making, in government and in the private sector, impacts that are often neglected to be mentioned in scientific evaluations. Finally, a measure of impact has to do with the generation of critical mass and the creation of new opportunities for Chilean science to develop. One of our strategic decisions has been to form a network of researchers that can use the expertise available at the CGR for their own independent projects. We have directly collaborated with numerous groups and scientists making possible that their projects advance far more robustly and efficiently. This “spillover” effect was also intended from the outset, as we were aware that our center offered the opportunity to disseminate an expertise and know how that many Chilean scientists were in need of.

I will summarize the salient results and effects that the CGR has generated in the first five years of operation in the following paragraphs emphasizing the qualitative outcome of the project, considering that details on the products themselves will be forthcoming in the appropriate sections of the report.

**Strategic achievements**

Of the few projects involving genomics carried out before 2010 in Chile, none had achieved results that could be considered to significantly improve the capacities or make resources widely available for work in this area. This was mostly due to the modest funding these projects received and their limited scope. The FONDAP award received by the CRG thus constitutes a turning point, since the project involved a critical mass of researchers from three different universities and funding that, for Chilean standards, was sufficient to generate an ambitious plan. Simultaneously -and critically for the success of this endeavour- two awards were given for projects in infrastructure: a next generation sequencing platform (later to become OMICs Solutions) and a very large computing facility (eventually, the NLHPC), funded by the CONICYT Program for Major Scientific Equipment. This fortuitous conjunction of events, allowed us to forgo the need for investing in equipment and to focus on the experimental aspects of our project. Today, we can say that, in our country, we have all that is needed for research in advanced genome science: a Center with the capacity to identify relevant questions and the know how for addressing them, the latest sequencing technology, and the most advanced computing power in the region. This fulfills our first, and most important, strategic goal.

As is the case with many Centers of Excellence in the country, the CGR is made up of investigators that are geographically and institutionally separated. Thus, there is no physical building or facility that groups us, a fact that we have sought to compensate in different ways. As a second strategic goal, we had wished to ameliorate this problem and we have attempted to find a solution albeit without a positive result to this day. Nonetheless, we have still managed to forge an identity as a Center and to become visible to the national and international scientific community (the third strategic goal). We have been approached by colleagues for collaborative projects, by the press for opinion and commentary and by government institutions for assistance and policy making. Besides these measures of recognition, scientists have started to join the CRG based on its reputation, in addition to the usual mechanism of recruitment that has to do with the visibility of its scientists as individuals. Three young Chilean researchers that trained as postdocs abroad and that had achieved independence, have returned to the country and have started projects within the CGR. All three are now seeking positions in academia (one successfully as of this writing) but they were able to return to the country only because the CGR gave them space, funding and an appropriate atmosphere for their re-insertion. In addition, five of our own postdocs were able to obtain academic positions in Chilean universities. Of these eight starting young scientists, two have been incorporated as Associate Investigators of the CGR. We've also hosted numerous visiting professors and one sabbatical stay at the Center, all of which speaks to the attractiveness of the CGR for career development.

**Scientific achievements**.

A major turning point in the history of the CGR was the appointment of an internal Scientific Advisory Committee that could guide us in identifying research directions that would be most fruitful with our resources and capacities. Reaching a definition of exactly what we should do as a Center was not easy and, admittedly, was not achieved until late in the second year of our funding period. The CGR could not simply be a "genomics center", dedicated to genome sequencing for the sake of it. Our expertise and history has been in fundamental biological research in cell biology, molecular genetics and regulatory modeling. However, our backgrounds in different organisms (microbes, plants and animal systems) and approaches (metabolic networks, cell differentiation, developmental genetics, etc.) meant that it was not trivial to find common ground for a collective project. It is tempting, and is commonplace in centers without a clear focus, to continue along the research lines that have been successful previously, and which are likely to continue to produce a strong publication record. Our Advisory Committee was key in suggesting that we look for Center Projects, that involved at least the group of six PIs and ideally all of the AIs and that would make us stand out with a clear identity (not just as the sum of the parts). We had

developed the idea of the relationship between environment and genome structure and function in our original proposal but had failed to define exactly where to examine this broad question. The concept of the "Natural Laboratory" put forth by the Chilean scientific community to stimulate research on national problems and resources was the ideal fit to our question.

We chose three main areas of inquiry which were transformed into the principal scientific goals of the project: extreme genomes, relevant genomes and gene expression in cells. The first objective, extreme genomes, has precisely as a feature that it aims to discover how environmental stress has shaped the genomic landscape of diverse groups of organisms, both in the evolutionary context and in terms of regulatory mechanisms. The site chosen for this analysis is the high altitude steppes of the Atacama Desert located in the North of Chile, one of the driest areas on earth. Plants and microorganisms that live in this environment have been collected and have undergone genomic and transcriptomic analysis, providing the first metagenomic analysis carried out in an entire ecosystem that includes species from different phyla and includes multicellular organisms. We have added animals as well: fish living in salt lakes and amphibians living in thermal springs have been analyzed in parallel. Two additional species have been examined, that come from different environments, but that have a special developmental mechanism to cope with stress and long periods of stasis: annual fish of the genus *Austrolebias* and the sporadic flowering desert plant *Cisthante longiscapa*. The work carried out within this objective has advanced and is at a stage in which the initial series of publications is near submission. Papers are under review for the work on amphibians (*Rhinella* frogs), *Orestias* fish and are in preparation for plants and microbes.

The second objective, relevant genomes, involves a group of projects that are of high national interest and are natural areas in which a Center such as the CGR can have a very high societal and economic impact. Firstly, we embarked on a Chilean Human Genome Project, as the peoples of our country have not been analyzed from a genomic point of view. This is especially true of our indigenous nations, which have admixed extensively with the European colonizers to generate an interbreed that makes up the large majority of the Chilean population. In order to detect genetic variants specific to this group of humans and to advance our ability to understand the genetic basis of prevalent diseases, we fully sequenced twelve individuals from the *Huilliche* race. The results show that there is a large cadre of new markers that can be made available to the human genetics community for association studies and there are also new insights as to the history of the original peoples of our country and continent. Particularly insightful has been the identification of markers linked to metabolic diseases, that are especially frequent in Chile and are related with diet. This work is one of the main scientific achievements of the CGR in this period and we have sought to publish the first set of results in a top tier journal. This process is still ongoing as of mid 2015 and the review cycles in three journals have been extensive, but we can say that the data is solid, relevant and will have international impact. The other relevant genomes that we have worked on are of species that have economic interest. One is a microbe that infects cultivated fish and causes huge losses to the aquaculture industry, *Piscirickettsia salmonis;* it has a very repetitive genome (unusual in prokaryotes) making it extremely challenging to assemble. We also sequenced and assembled the genomes of a set of bioleaching bacteria that are critical for the process of metal ore lixiviation; this project was carried out jointly with a biotechnology company that is a subsidiary of the national copper corporation. Furthermore, we sequenced the genomes of two plants, crops that represent a large fraction of chilean fruit exports, the table grape and the peach. And lastly, the Atlantic salmon, where we have collaborated with an international consortium to sequence its genome. In the latter case, our bioinformatics group was able to contribute significantly to the assembly of the salmon genome, a task made exceedingly difficult by its tetraploid nature and the high amount of repetitive DNA it contains. Articles on these genomes have been published or are in the final stages of preparation and we consider them as finalized projects.

The third main scientific objective encompasses a set of subtopics that were grouped collectively under the name Gene Expression in Cells. This objective includes all of our work investigating molecular mechanisms of gene regulation and epigenetics, the control of cell differentiation and behaviors (migration, differentiation, metabolism, growth, senescence) and the response to stress. We also include here the research in theoretical biology and network modeling, as related to the biological systems under study. Within this third objective are numerous specific lines of inquiry that have represented the bulk of our scientific productivity during the first four and a half years. As these questions require experimental approaches, we of course use model laboratory systems (mouse, arabidopsis, zebrafish, fly, mammalian cell lines, etc.). While these are areas of research that we will by no means abandon, they will turn increasingly towards the organisms we have analyzed in objectives 1 and 2. For instance, we are examining the epigenetic mechanisms and regulatory small RNAs that could explain environmental adaptation in the amphibian *Rhinella* and in *Orestias* and *Austrolebias* fish. Or, we have found specific metabolic networks present in many taxa of plants inhabiting the desert environment, like exacerbated nitrogen capture mechanisms or moisture retention structures in their leaves. As genome editing technologies are no longer limited to laboratory model organisms (thanks to CRISPRs, for instance), we foresee the ability to test hypotheses directly in the species of interest (see proposal for continuity of the Center).

**Synergy**

One of the most difficult aspects to deal with in a large, multidisciplinary Center of Excellence, is the consolidation of a collective spirit that motivates researchers to contribute to common goals. Science tends to recognize individual achievements while consortia that group meritorious scientists is faced with the challenge of making them work together sacrificing personal gain. We feel that at the CGR we have achieved a productive balance protecting the individual interest yet creating synergies between its members. There are several factors that have made this possible. First, we have distributed funds specifically aimed at supporting Center projects, both in terms of expendables and personnel. Second, we have expanded our use of the "Jamboree" (the meetings where all participants of a project meet for brainstorming, results discussion and manuscript preparation) for all Center projects. Thirdly, all PIs participate directly in all Center projects; authorship in the papers requires involvement at all possible stages of project fulfillment. All of these measures have driven the work forward at a higher speed than was the case before them, and has ensured a high degree of commitment by all involved. Again, the measurable impact of this collective effort will be seen in the increase in co-authored publications and co-mentored students and postdocs. We have only just begun to see this effect as the articles describing these projects are starting to appear in this fifth year of the project; they are likely to be even more significant during the second five-year period of the CGR.

**Training**

The number of students and postdocs that have worked at the CGR labs is one of the strongest measurable outputs that we can exhibit at this point in our development as a Center. This speaks volumes about our performance in two important aspects: CGR group leaders are at the most productive stages in their careers thus being able to easily recruit talent; and young scientists are inspired by the projects available at the Center. In a competitive job market, our trainees are able to find positions, which also attests to the attractiveness of the topics we encompass and the need for this type of researcher. How good are our students and postdocs? Most of them are able to obtain outside funding (PhD and postdoctoral fellowships) and they come from the best universities in the country or from prestigious foreign institutions. As a policy, we encourage them to spend time abroad, in the labs or centers of our collaborators, a fundamental step in their development. We also bring in outstanding colleagues to speak at conferences and symposia and support CGR trainee participation in national and international

meetings and courses. They have also had the chance to organize events themselves, which they have taken advantage of.

**Collaboration networks**

One of the strategic objectives of the CGR was to generate networks with other centers and to attract scientists both locally and internationally. We have made substantial progress on both fronts. At the national level, we have strong interactions with centers and universities all over the country. Our most important partner is the Center for Mathematical Modeling at the University of Chile as they are providers of the critical computing infrastructure and expertise required for genome scientists. We also have strong ties with the National Genomics Resource Center (named above), that has installed the most relevant sequencing capacities in the country. Internationally, we have established ties (collaboration agreements and/or joint projects with funding) with institutions that include the Australian Regenerative Medicine Institute at Monash University and the Center for Organismal Studies at the University of Heidelberg. We have hosted several visiting scholars for short or long stays at the CGR and a large fraction of our postdoctoral fellows are foreign. We have an incipient relationship with a new FONDAP Center, the Advanced Center for Chronic Diseases (ACCDis), with whom we are planning an ambitious human genomics project in the second term of the CGR (see proposal).

**Interaction with the Private sector and Government**

We have always claimed that genome science should have a strong impact on the productive sector. Many of our projects are related to species with critical economic impact in the country. For instance, our work with biomining bacteria, together with BioSigma, a Chilean biotech company, has generated important knowledge for bioleaching, a process aimed at increasing efficiency in the extraction of copper ore. About X percent of copper production is generated by this means, representing a huge return on investment for the copper industry. Likewise, we have supported the fruit industry consortium with our work on table grapes and peach, two of our main exports in this sector of the economy. Finally, the aquaculture industry has worked together with the CGR on salmonid genomics and on research involving one of the critical pathogenic bacteria that generates losses for the industry: *Piscirickettsia salmonis.*

The Chilean Human Genome Project, which we are leading, will provide an unprecedented resource for the country, both for public policy and for health care providers. We expect that our current data, together with the accumulating genetic information we are generating on the Chilean population will be incorporated into a national platform that will be freely available. The Chilean Minister of Health has already expressed interest in this project and is likely to support this effort. We have also been approached by clinicians and hospitals that want to set up their own genomic resources and analysis centers. As is happening all over the world, there is a drive to implement the personalized medicine model, in which analysis of the genome shall become a key item of information for clinical decisions. Interestingly, this is an area where scientists trained at the CGR can expect to find job opportunities, a desirable expansion of the market for our professionals. This has already happened in one of the research areas that we have carried out at the CGR: numerous young scientists trained in stem cell research at our center, have been hired by private clinics which are also implementing cell or tissue banks and stem cell-based treatments.

Lastly, we have recognized a new area of influence of the CGR: environmental protection. Our work in the north of Chile examining species that can tolerate high concentrations of heavy metals and other pollutants has attracted the attention of environmental health authorities. We are generating an agreement of cooperation with the head of environmental protection in the Antofagasta Region aimed at exploring the use of microorganisms to bioremediate contaminated soils in the area. This problem has become exacerbated with unexpectedly damaging rainstorms which overwhelmed many of the leaching

pools that the mining industry has there spilling their contents into inhabited areas. Our more traditional work with zebrafish and other aquatic organisms has also led us to recently publish two environmental regulation guidelines (one of them an adaptation of an ISO standard used in Europe for environmental monitoring). These guidelines can now legally be applied in Chile by industry and government to ensure that continental waters are adequate for human consumption or other uses.

**Outreach and dissemination of results**

Over the past four years, our scientists and our research have appeared prominently in national media. The sequencing projects have made the news on several occasions as well as individual achievements by our investigators. Many of these instances are summarized in the appropriate section of the report. We can attribute these reports as the result of the efforts of our journalist who was hired for that purpose. She has also been in charge of our web page, press releases, appearances on radio shows and dissemination of events to the public and to the scientific community. Our very successful seminar series, the "Interactomics", were very well attended as were conferences and symposia organized by the CGR. Every year we funded and organized at least three activities with international visitors which maintained a high visibility of our Center with colleagues. We have a very recognizable "corporate image" with all of the elements that are necessary to become widely known.

We invested strongly in outreach to the general public, an important task for our center. Some products of this effort are videos describing our research and that of other institutions as a way to approximate science to the layman. All of us are heavily involved in secondary school teacher training and we have had numerous courses for teachers and demonstration labs for children. In late 2015, we are organizing a workshop to design a portable lab that can take our science to the classroom, an undertaking we expect to carry out in 2016.

**Infrastructure**

As stated above and in previous reports, it is the desire of our investigators to eventually coalesce into a physical building that can house are labs and facilities. Of course, this aspiration is beyond the reach of our FONDAP budget and it depends on our ability to obtain specific funding for infrastructure. It also requires a careful negotiation with our mother institutions (universities) so that they will agree to such a multi-institutional arrangement. Our efforts have not been successful thus far although authorities that we have met with (government, CONICYT and universities) are all enthusiastic about the prospect of creating an appropriate infrastructure and location for the Centers of Excellence the country has. We will continue to pursue this goal in the second stage of our Center's funding.

The above notwithstanding, we have been able to rent offices and work space to house our administrative staff and a team of bioinformaticians. This space also has a large meeting room where we have held our "Jamborees". Additionally, we have made improvements to facilities in our labs in the different campuses that have made an impact on our research capabilities. Rooms for genomic work housing sophisticated equipment have been made. Also, we have constructed facilities for animals and plants (a new zebrafish facility, a mouse facility, human stem cell lab and housing for *Arabidopsis* and other plants have been built with CGR contribution).

## III. RESULTS:

### 3.1.- Research

### Research Line 1. Extreme Genomes

## 1.A. Plants

### *De novo* transcriptome sequencing of Atacama plant species to understand molecular determinants of adaptation to a hyperarid environment.

The Atacama Desert is one of the oldest and driest places on Earth. Organisms living here are exposed to extreme environmental conditions that include extremely low water availability, high diurnal temperature oscillations, nutrient-poor soils, and high levels of solar radiation. Despite these harsh conditions, the Atacama hosts a surprising diversity of animal and plant life. In the central Atacama Desert, the western slopes of the Andes provide a natural altitudinal gradient of environmental parameters, such as rainfall and temperature. As a consequence, various plant communities succeed each other at different elevations supporting animal life: 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.). In this aim, we wished to gain insight regarding molecular mechanisms that explain plant diversity on the western slopes of the Andes in Atacama, and reveal the genetic bases of adaptation to this extreme environment.

To achieve these goals, we used *de novo* transcriptome sequencing to characterize the transcribed fraction of the genome for the community of plant species found in this environment. We selected an altitudinal transect on the western slopes of the Andes in the Atacama between 2400 and 4500 m.a.s.l. which spans the limit for life in desertic conditions (no plant life below 2400 m.a.s.l. due to lack of water) and the limit for life in altitude (no plant life observed above 4500 m.a.s.l. due to extreme low temperatures). We collected samples of all plant species present in this transect throughout five consecutive years. We found 64 species belonging to 51 genuses and 20 plant families, fourteen of which are endemic families (Table 1). Four plant families account for more than half of the species: Asteraceae with 14 species (22%), Poaceae and Fabaceae with 8 species each (13%), and the Solanaceae with 6 species (8%). Ten families are represented by only one species. Plant families from the present Atacama/Andes transect derived from diverse plant families, some of which diverged almost three hundred million years ago (for example, *Ephedraceae*). Four transect plant families with more than 1 species showed enrichment as compared to worldwide plant richness (pvalue<0.05): Montiaceae, Verbenaceae, Boraginaceae and Solanaceae. Boraginaceae was also found enriched in another altitudinal transect Saudi Arabia (Masrahi et al (2011)) Two plant families are enriched when compared to the Atacama transect richness to Chilean flora richness: Krameriaceae and Solanaceae, and one is depleted, the Asteraceae. These results indicate that species richness in Atacama does not mirror worldwide, arid altitudinal environments, or Chilean flora, and that some plant families have been more successful in the extreme conditions of the Atacama/Andes transect.

We sequenced 32 out of the 64 plant species found (bold in Table 1), representative of the phylogenetic diversity in the transect, covering more than 90% of the biomass and including plant clades containing species of high agricultural value (e.g. *Fabales, Poales,* and *Solanales*). Sequencing was carried out using Illumina paired-end technology. Predicted proteome size and the enzyme/non-enzyme ratio were similar to the genome of six reference model plant species. The largest predicted proteome found belongs to the diploid (2n=22;

Christopher and Abraham, 1971) Pre-puna/Puna annual C4 grass *Aristida adscensionis*, with 44,355 predicted proteins.

To better understand the conditions to which plants in Atacama are exposed, we carried out a complete elemental and chemical analysis of the soil in all stations sampled. Soil samples analyzed showed characteristics of desert environments, with more than 80% sand content and fairly low nutrient concentrations. Most macro and micronutrient concentrations, pH and conductivity measured in the soluble soil fraction showed marked altitudinal variation. The exception was total nitrogen, organic matter, nitrate and ammonia levels which were found low regardless of altitude. Species richness positively correlated with organic matter and total nitrogen. This result suggests that in addition to water and temperature, soil N availability is a major determinant for plant life in these environments. Consistent with this hypothesis, the deduced proteome for most species sequenced (except Fabaceae and Solanaceae) displayed a higher C:N ratio as compared to the proteome of sequenced reference species, indicating a possible shared mechanism of nitrogen economy for most plant species in the transect.

Genome/transcriptome pathway-databases are a powerful resource to explore plant metabolism from non-model plant species. Pathway-databases were generated for each Atacama species and compared to reference species. Over half of the pathways predicted for any species (363 out of 692, 52.5%) were found in 95% of the species. When compared to PMN Universal Plant Pathways subset version 5.0 (Zhang et al., 2010), 82.1% of the pathways (151 out of 184) in our seed plants pathway dataset match the PMN Universal Plant Pathways subset. As expected, shared pathways are mainly involved in primary metabolism. Unique pathways at the family level were found for Fabaceae (18 out of 692, 2.6%), Solanaceae (6 out of 692, 0.9%), Poaceae (4 out of 692, 0.6%) and Asteraceae (1 out of 692, 0.6%). These pathways were involved in specialized metabolism, mostly biosynthesis of compounds typical of each family such as isoflavonols for Fabaceae, the phytoalexin capsidiol in Solanaceae and cyanidin dimalonylglucosides in Asteraceae. Exceptions to this rule include pathways PWY-822 and PWY-842 found in Poaceae, which are involved in carbohydrate metabolism, and pathway PWY-5373 involved in biosynthesis of fatty acid calendate, known to accumulate in high amounts in the Asteraceae plant *Calendula officinalis* (Qiu et al., 2001). Expansion of gene complement size was detected for Solanaceae and Poaceae (Figure 1). Gene expansion may result from selective pressure in Atacama and suggest mechanisms of adaptation. For example, Solanaceae found in Atacama displayed twice the number (t-test, $p<0.000$) of stelar K+ outward rectifying channel (SKOR) homologues as compared to the *A. thaliana* genome or reference species from Solanaceae. It has been postulated that this channel might be involved in adaptation to drought conditions through osmotic adjustment of the root cells, which can be accomplished by mediating a reduction of K+ translocation to shoots (Figure 1, Annex).

Gene Ontology enrichment analysis found a substantial overrepresentation of transcripts related to abiotic stress tolerance such as response to water deprivation, cold stress, and oxidative stress. This result suggests plant species in Atacama share and overexpress known mechanisms of plant abiotic stress tolerance studied in model species. In order to gain deeper insight and identify specific biological processes that could explain

adaptation to this desert environment, we used BigPlant to carry out a functional phylogenomic analysis on the transcriptome data (Lee et al 2011). The strength of the BigPlant phylogenomic pipeline lies in its ability to discover evolutionary signal even in the absence of complete genomes. In fact, the initial applications of this phylogenomic pipeline included only 5 fully-sequenced genomes and 145 species represented by incomplete transcriptomes (Lee et al 2011). At each branch point in the phylogenetic tree, composed of thousands of orthologs, the phylogenomic matrix identified the contribution (e.g. positive branch support, PBS) that each gene provides to species divergence. By identifying genes that support divergence of species sharing a common trait, this approach was able to determine trait-to-gene relationships (Lee et al 2011). Further, GO-term analysis of genes that provide positive branch support (PBS) was then used to identify the biological processes underlying evolution of the trait. When applied to our transcriptome data set, we found three biological processes over-represented in branches that separate many Atacama species from reference species: N-metabolism, seed development and RNA metabolism. This analysis indicates biological processes that are likely key for adaptation of Atacama species to these extreme environments and provide new insights into mechanisms for evolution of plant abiotic stress tolerance.

**Genomics, Transcriptomics and Metabolomics 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. This indicates that these plants have certain genetic features, which allow them to germinate and stay alive for weeks under minimal conditions, and also to achieve dormancy, for years if conditions are not met for germination. One of the dominant species is *Cistanthe longiscapa*, a very abundant plant in the Blooming Desert, which grows even if the amount of rainfall is not enough to trigger this phenomenon in other species. It is of great biological interest to identify the genes that participate in regulation of these unique features. We have performed a transcriptomic and genomic sequencing effort aiming to identify genes that may be involved in stress tolerance, activation of germination and developmental stasis, the salient and unique features that this plant possesses. This project is a collaboration of the CGR with Andrés Zurita from INIA-Intihuasi, and Francisca Blanco, Claudio Meneses and Reinaldo Campos of UNAB. In addition, we have started a metabolomics characterization of some the organs of this plant through a collaboration with Dr. Fernie from the Max Planck Institute for Plant & Molecular Physiology .

A first draft of the genome, at 54X coverage, indicated a genome size around 500 Mb, similar to the size estimated by flow cytometry (around 600 Mb). In addition, Kmer analyses showed that *Cistanthe longiscapa* is diploid. Interestingly, the SNP rate is very high (1/53) and there is no evidence of ancestral duplication of the genome. These results suggest that a

highly polymorphic genome could help to cope with the harsh conditions to which this plant lives in. Regarding the repertoire of genes, around 30,000 genes were found and 80% of them were supported by RNA-seq data. Currently we are performing the annotation and the analyses of the differential expression of genes among organs.

The metabolomic analysis, along with the identification of genes by genomics and transcriptomics, is helping us to map metabolic networks playing important roles in the homeostasis of *Cistanthe longiscapa*. For instance, we were able to identify raffinose as well as the genes involved in the metabolic pathway leading to the formation of this carbohydrate, which is highly accumulated in the plant stems compared to flower tissues. Interestingly, the raffinose family oligosaccharides are characterized as compatible solutes involved in stress tolerance defence mechanisms. Then, this kind of approach is helping us to reveal mechanisms that can be used by plants living under extreme conditions.

**B. Microbes.**

**Analysis of soil bacterial composition in desert plants: testing the bacterial recruitment hypothesis.**

In this project examining soil biodiversity, we seek to study the potential modification of bacterial communities growing in soils located close to plants (around the roots) and soil found at a short distance away with no evidence of recent presence of plants. Taking advantage of the unique environmental scenario present in the Atacama Desert, the area of study selected (23°50'S, 67°70'W, North of Chile) considered an altitudinal gradient transect which includes three different ecosystems: Pre-Puna (2400-3300 m.a.s.l.), Puna (3300-4000 m.a.s.l.) and Steppe (>4000 m.a.s.l.), whose classification is due to the composition of plants present in the area (n=40-60). Until now, we have quantified diversity; relative abundance and taxonomic composition of bacteria associated with 14 plant species distributed along this altitudinal transect to evaluate the role of edaphic and biological factors on soil microbial communities (Figures 2 and 3). We expect that functional activity and/or taxonomic structure of the microorganism community will be differentially affected by plant species-specific metabolic features.

**C. Animals.**

**Local adaptation detected in a population of high altitude amphibians**

In amphibians, growth and differentiation depend strongly on the environment. Although it has been documented that, in the Andean toad *Rhinella spinulosa*, water temperature strongly influences the life cycle and morphology, the specific genes involved in these adaptive responses remain unidentified. In this study, we document the genome-wide transcriptional changes induced by thermal variation in *R. spinulosa* and screen for differentially expressed (DE) genes possibly involved in local adaptation to temperature (See figures in Annex, Pastenes et al). A common garden experiment was designed with larvae obtained from three different habitats: El Tatio, a geothermal stream, and Catarpe and Farellones, ponds with seasonal and daily variable temperature. Larvae from these localities

at two different developmental stages, Gosner stages 36 and 42, were experimentally exposed to 20 and 25 ºC. An RNA-Seq trial with 12 samples, *de novo* transcriptome assembly and annotation was performed. Because of the strong genotypic differences evidenced by Farellones, this population was excluded from the transcriptomic comparisons. Bioinformatic analysis revealed 6,241 DE genes when the comparisons were by temperature variation. Catarpe showed a total of 6,220 DE genes while El Tatio showed only 21. Analyses of the data detected one gene with increased expression in samples from El Tatio maintained at 20 ºC compared to its natural condition of 25°C. This gene codes for a pre-prohormone, pro-opiomelanocortin, a protein tightly related with metamorphosis. Together, these results suggest that the lack of differential expression of genes in the El Tatio population may have a relevant role during adaptation of *R. spinulosa* to unchanging environmental thermal conditions. We have submitted a revised version of a manuscript describing these findings to the journal Molecular Ecology.

**Genome evolution of fish of the genus *Orestias*, *Cyprinidonitformes***

The cyprinidontiforms 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 remarkable 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 identified as one of the species that we wanted to characterize at the genome level since the beginning of our Center. Now, we have added three more species to the list of target organisms as we intend to compare the genomes of these recently speciated groups. To date, we have a complete genomic sequence of *O. ascotanensis* and a draft for *O. gloriae*. The quality of the data we have obtained from the genomic DNA prepared from single individuals is outstanding and shows that this species, unlike other teleosts sequenced thus far, has a genome that has less repetitive elements than other cyprinidontiforms, which normally make assembly exceedingly difficult (i.e., Atlantic salmon or *A. charrua*, see below). The genome of *Orestias* fish is less than 1Gb. To annotate the genome, we used predictor software and RNASeq data; we find around 21,000 genes in these species which is in the expected range. We have also examined the miRNAs present in this organism (predicted and experimentally validated) and find that the number of families is much lower than in other vertebrates.

For *Orestias ascotanensis* genome we use a combination of 4 types of DNA libraries with insert ranges from 280 to 10k pair bases, generating nearly of 1,200 million of reads. Genome size of *O. ascotanensis* has been reported as 756 Mb based on Kmer analysis and our procedure to assembly the raw data allow as generate an scaffolded genome with a total length of 696.3 Mb (Table 1; see figures in Annex “Orestias”). Using CEGMA analysis for validate genome assembly we find 100% of the 248 ultra-conserved core eukaryotic genes, 98% of them as complete genes. We also prepare several RNAseq libraries from whole adult’s species in order to aid in annotation procedure. With this assembly we predict 33,485 transcripts which correspond to 21,034 coding genes. Using an annotation pipeline from the Broad Institute, we were able to annotate 19,552 (92%) of the coding genes (Table 2). Using

an orthology and clustering analysis between O. ascotanensis and another 14 sequenced and annotated genomes we can identify 20,411 groups of orthologs proteins composed by 273,389 proteins. Among these, 1,970 were detected in all organism analyzed and were used to phylogenic reconstruction. These allow as locate O. ascotanensis in a monophyletic group together with *Poecelia Formosa* and *Xiphophorus maculate*, two members of cyprinodontiformes family (Figure 1). Orthology analysis also identifies 2,119 groups which contains paralogs proteins including *O. ascotanensis* genome. From these, 109 groups have unique genes from all species except *O. ascotanensis*, which comprises a total of 263 of paralogs families exclusive for its genome. Preliminary analysis of these genes indicates functions, assigned by gene ontology terms, involved in metabolic, DNA repair and oxide-reduction process. Also we identified in the genome those genes which are under positive selection pressure. A total of 393 genes were selected under these conditions with distinct functions assigned by the different gene ontology hierarchies: metal binding, protein binding, DNA repair, transport and cytoskeleton organization, among them (Figure 2).

Also in the genome we identified 166 sequences of miRNA using the information available at mirBase. To confirm the prediction, we sequence a miRNA library with RNAs from the same sequenced individual. Once raw data were filtered and assembled, we can identify 489 putative sequences for miRNAs from which we can verify a 100% of identity in a 148 of the miRNA predicted in DNA sequence. Until now we are able to identify 270 putative interactions of the sequenced miRNAs in a 191 proteins of *O. ascotanensis*. We are comparing these data with another sequenced fishes and vertebrates in order to define if these ratios of miRNAs/targets are similar in other species. Until now we can define ontology definitions for the 191 groups and we founded that heterocyclic compound binding, anion/ion binding and development (Figure 3).

**The genome of *Austrolebias charrua*, an annual killifish**

Annual fish are freshwater teleosts found in South America and Africa that inhabit an extremely variable environment. They develop and reproduce in seasonal ponds that dry out during the summer. The survival of the species becomes entirely dependent upon buried and dried embryos, that hatch during the next rainy season and present a peculiar development, both features unique in vertebrates. Annual fish exhibit a unique developmental strategy in which embryogenesis occurs within a reaggregated mass of previously dispersed cells. On the other hand, they can enter a state of reversible developmental arrest (diapause) at three distinct embryonic stages. *Austrolebias charrua* is a Uruguayan annual killifish inhabiting the grasslands of the Rio de la Plata delta whose development has these features. Our aim of is to analyze the genomic structure and control of gene expression during diapause in *A. charrua* assuming that its adaptation to a changing environment has evolved by the acquisition of genetic traits that sustain developmental and physiological plasticity.

From two paired-end libraries (HiSeq2500,150bp) with different size of insert selection (280bp, 15kb) we assembly a first draft of the *A. charrua* genome, which, using Kmer analysis, results in an estimated genome size of 2.942 Gb, with a coverage of 11X. Parallel to this approach, we sequenced transcripts of different developmental stages and tissues using paired-end libraries. This allowed us to assemble ~ 114000 transcripts (46,800 genes) with a

mean back-mapping of 87%. For these genes we are able to predict 22478 protein sequences and using several databases (BLAST, Swissprot, Trembl and Kegg) we asigned annotation for 21867 genes.

The information obtained from transcriptome assemblies was used for genome annotation. The consensus transcriptome of the developmental stages of *A. charrua* (64445 transcripts) was used to improve the genome asembly, allowed us to increase the number of scaffolded contigs from 8% to 13.3% and to diminish the average length of break size from 2200 to 738. CEGMA software was used to assess the completeness of the genome assembly by evaluating the presence and completeness of a widely conserved set of 248 eukaryotic proteins. After use RNAseq data for help in genome assembly we identified a 61% of full-length 248 core proteins.

The analysis of the transcriptome and genome allow us to almost completely ensemble and partially annotate the *A. charrua* genome. In this process we determined that 68.1% of the genome becomes masked as repetitive sequences. To circumvent the problems derived form its repetitive nature, and evidenced in this first assembly, we are planning to re-sequence a genomic library using a PacBio technology for large library sizes (~ 15kB), which will permit the better identification and location of repetitive elements (RE) along the genome.

From our analysis, and to our knowledge, *A. charrua* is the vertebrate with the highest amount of RE, even higher than that of the Atlantic salmon, *S. salar*. Interestingly, the proportion of the different elements is exceptional among teleost, with LINE elements being the most prominent. From a genome expression perspective, this offers a potentiality for the emergence of gene regulatory motifs or building blocks for genome structural features selected in this animal. On the other hand, and from a population genetic point of view, this might reinforce the notion that genome complexity arises as the resultant of evolutionary drift of genomic variations produced by the small population sizes in *A. charrua*. Our preliminary study suggest that an extreme habitat does impose a pressure to select for complex genomes or annual life cycle since *N. furzeri*, an African annual fish inhabiting a similar environment, has a genome of only a third of the size of *A. charrua*. Similarly, genome complexity does not appear to be directly related with diapause, another feature of *A. charrua*, since phylogenetic studies have suggested convergent evolution for this trait.

We are currently analyzing the position of the REs to examine their gene regulatory potential. In the next period this *in silico* approach will be complemented with transcriptome data from pre-, post- and diapause animals to build up an understanding on the conserved and divergent mechanisms operating in this extraordinary process.

## Research Line 2. Relevant Genomes

### A. The Chilean Human Genome

**The Mapuche-Huilliche Genome Reveals Links to Prevalent Diseases in Native Americans (for Figures, see manuscript by Vidal et al in the Annex)**

Sequencing complete human genomes has greatly expanded knowledge of our genetic diversity, providing insights into the evolutionary history of man and the bases of

human diseases. Large-scale genomic initiatives such as HapMap (Altshuler et al., 2010) or the 1,000 Genomes Project (Consortium, 2012) have shown that each individual possesses millions of genetic variants. This genomic information and genome-wide association studies (GWAS) are powerful approaches to identify genetic variants that may influence susceptibility to develop common diseases. While current high-coverage full genome efforts have mostly focused on Old World continental groups (Europeans, Asians and Africans), there is still limited information concerning the genetic structure of ancestral American groups. Addressing genetic variation in Native populations represents a first step towards understanding the molecular basis for common diseases affecting modern Native American and Latin American admixed populations.

We sequenced at high coverage the complete genome of 11 individuals belonging to a native Mapuche-Huilliche population from Southern Chile (HUI), which is considered to be one of the original populations who settled the Southern Cone of South America in the pre-Columbian era. HUI individuals sequenced correspond to a community living in Huapi island, Ranco Lake (latitude 40°13'27.62"S, longitude 72°22'50.16"W). DNA samples were sequenced using the combinatorial probe-anchor ligation and DNA nanoarray technology of Complete Genomics (Drmanac et al., 2010). We obtained an average of 85% genomic coverage of at least 30X and an average of 98% exomic coverage of at least 30X, with 96-97% high-confidence calls (Figure 1-figure supplement 1b). We identified 464,952 (7.9%) SNVs which are not included in dbSNP build 138 release or do not have a reported frequency in the 1,000 Genomes Project phase 1 (Consortium, 2012) database (Figure 1-figure supplement 3). These SNVs were catalogued as “novel”. Likewise, analysis of genetic variants indicated that 270,640 (66.5%) insertions and 61,780 (14.6%) deletions are novel and observed in at least 1 HUI individual (Figure 1-figure supplement 2b, 2c and 3).

Mapuches are the modern representatives of one of the most prominent indigenous groups in the Southern Cone of America. Mapuche people descend from early hunter-gatherers who colonized the subcontinent about 15,000 years ago (de Saint Pierre et al., 2012). We used the ADMIXTURE (Loh et al., 2013) software to determine individual ancestries of the sequenced HUI, comparing a set of SNVs to those of ancestral founder populations including Yoruba from Ibadan in Africa (YRI), Chinese Han from Beijing (CHB) and Utah residents with European ancestry (CEU) (Consortium, 2012) (Figure 2a). The ADMIXTURE model indicates that most of the genetic contribution in the HUI genomes sequenced comes from their own Native American ancestry, with negligible contributions of Asian, European and African ancestries. Moreover, when including sequence data from an American admixed population with Mexican Ancestry (MXL) from Los Angeles, USA (Consortium, 2012) in the analysis, the HUI population behaves as a founder population for admixed Mexicans, contributing with an average of 32% of their genetic composition. This result indicates that HUI individuals sequenced are representatives of an ancient Native American population (Reich et al., 2012, Silva-Zolezzi et al., 2009, Zhou et al., 2013). Moreover, when we analyzed the population structure using PCA analysis (EIGENSTRAT) with data from Native American populations from the Southern Cone, we found that HUI are situated among other Chilean populations, including non-insular Huilliches (Reich et al., 2012) and are close to admixed Chileans from Santiago, Chile This result indicates HUI individuals

sequenced, from an isolated community of insular origin, do not exhibit genetic drift and is consistent with sequence data representative of American indigenous populations. Finally, analysis of mitochondrial DNA showed that all HUI individuals belong to the Native American haplogroups C and D, two of the major pan-continental founder.

These results indicate the sequenced HUI genomes represent a valuable resource to explore genetic variation in Native Americans and its potential phenotypic impact. Moreover, these genomes will be key to identify markers that are specific for individuals with Native American ancestry for future studies.

Populations with different ancestries exhibit substantial genetic variation (Consortium, 2012). In order to identify highly divergent genetic variants in HUI as compared to other world populations, we calculated fixation index (Fst) values for SNVs with a calling rate of 90% or higher. Pairwise Fst values were calculated as described previously (Reynolds et al., 1983, Brinkworth et al., 2014) between HUI and populations from Africa (AFR: LHK, YRI and LWK), Europe (EUR: CEU, IBS, GBR, FIN, TSI), America (AMR: CLM, PUR, MXL) and Asia (ASN: CHB, CHS, JPT) (Consortium, 2012). High divergence Fst thresholds were estimated for each comparison using a genome-wide empirical distribution of Fst values based on 10,000 permutation tests and using a 95-percentile cutoff. Novel SNVs were also included when found in 3 or more HUI individuals. We evaluated the potential functional impact of highly divergent SNVs using knowledge present in the Genome-Wide Association Study (GWAS) Catalog (Welter et al., 2014) and other databases and predictive algorithms. Briefly, highly divergent SNVs were used to query the GWAS catalog to evaluate known associations with diseases or traits. In addition, SNPEff (Cingolani et al., 2012) and dbNSFP (Liu et al., 2013) were used to assess the functional impact of the variant on protein coding genes. Highly divergent SNVs with an associated functional impact were selected and termed Variants with Potential Functional Impact (VPFIs). We obtained a total of 937 VPFIs that map to 890 genes. This unsupervised analysis identified known variants in genes linked to common diseases in populations with Native American Ancestry such as type 2 diabetes (T2D), rs75493593 in the SLC16A11 gene and rs17584499 in the PTPRD gene (Below et al., 2011, Consortium, 2014) (Figure 3-Additional Data 2).

Genetic variation explaining differential susceptibility to disease or metabolic conditions derives mostly from studies in admixed Latino populations such as the T2D example mentioned above. Native Americans, particularly those from the Southern Cone, are a neglected group in population-based epidemiological studies. As a first step to prioritize VPFIs and understand the potential contribution of corresponding genes for common diseases or disorders in Native Americans, we analyzed overrepresentation of genes linked to diseases or disorders using available information on DisGeNET database (Piñero et al., 2015). We selected diseases with overrepresented genes in HUI as compared to at least 9 other populations (Figure 3). In order to visualize the 22 found diseases/disorders and 126 associated genes with VPFIs, we represented them as a network graph (Figure 3). Nodes in the network represent genes (squares) or diseases/disorders (triangles). Edges that connect the nodes in the network indicate known association between a given gene and disease/disorder (dotted grey edge) or known functional interactions between genes (solid purple edge) obtained from GeneMania (Warde-Farley et al., 2010). Network layout was

based on a community cluster algorithm from the clusterMaker plugin of Cytoscape (Morris et al., 2011). Interestingly, clusters found belong to three main types of disease or disorders: (i) Metabolic and cardiovascular disorders, (ii) Neoplasms and (iii) Inflammatory and immunity processes (highlighted in Figure 3, Figure 3-Additional Data 3). Fifty percent (14 out of 28) of the genes contained in the "metabolic and cardiovascular disorders" sub-network are related to T2D. It has been well documented that Native Americans including HUI and Latino populations of Chile are prone to develop disorders related to glucose and insulin metabolism leading to insulin resistance and T2D (Celis-Morales et al., 2011). Similarly, recent data indicates that some Native Americans and Latin populations with Amerindian heritage exhibit a substantial predisposition to dyslipidemias and cardiovascular events (i.e. coronary heart disease and stroke) (LaRosa et al., 2005, Aguilar-Salinas et al., 2009). Among genes related to these diseases, we found two genes with novel SNVs, one on PPP1R17 (chromosome 7, position 31736608) and on PROC (chromosome 2, position 128186443). These genes have been associated with hypercholesterolemia (PPP1R17) or thrombosis (PROC) in Asian populations (Ono et al., 2003, Yin et al., 2014) and might have a similar function in HUI populations. In addition, other relevant genes included in this sub-network participate in lipid metabolic pathways including ABCA1, FNDC4, INHBE and FADS2 (Oram et al., 2001, Lattka et al., 2010, Hashimoto et al., 2006, Weissglas-Volkov et al., 2013). The second sub-network, termed "neoplasms" includes diseases that have been reported with increased prevalence in Native Americans, including hepatocellular (Jim et al., 2008, Suryaprasad et al., 2014), colorectal (Perdue et al., 2014) and kidney (Li et al., 2014) cancers (Figure 3). Chronic liver diseases leading to cirrhosis and hepatocellular carcinoma, independent of the etiological factor, are more prevalent and aggressive in Latinos with Amerindian ancestry as compared to non-Latino white populations (Carrion et al., 2011) . Interesting genes in this sub-network are CYP1A1 (rs1048943), AKR1B10 (rs3735042) and FASN (chromosome 17, position 80040521). These genes have been implicated in many types of neoplasms, including hepatocellular, gallbladder (He et al., 2014, Sharma et al., 2014) and gastric cancers (Lee et al., 2010, Yao et al., 2014) in Hispanics and Native Americans (Siegel et al., 2012, Arnold et al., 2014, Andia et al., 2008) and variants found here might represent risk factors for these diseases specifically in HUI or Chilean populations. The third sub-network, termed "inflammatory processes and immunity" includes diseases or disorders related to inflammatory, immune or infectious conditions. Amerindian populations, especially those of the middle and Southern Cone, experienced a strong selection after the arrival of the Spaniard conquerors during at least the first century of colonization by the introduction of infectious diseases brought from the Old World into pre-Columbian America. In some countries like Chile, an 80 to 90 % reduction in the Native population at the end of the first century of colonization most probably by epidemics of measles, smallpox, influenza and other infectious disease (Crosby, 1976, Bengoa, 2003) . Interestingly, this sub-network contains most of the genes affected by novel variants (6 out of 9 genes with novel variants). It also contains half of the genes affected by variants divergent in all 14 populations as compared with HUI, indicating immune/inflammatory and infectious diseases in HUI might be partially explained by variants that are specific to this population. We hypothesize that VPFIs on genes related to immune/inflammatory may not necessarily reflect a differential predisposition

of HUI to these traits but that they may have inherited a peculiar constitutive profile with advantages or disadvantages for environmental-genetic interactions in contemporary lifestyle. Given the pleiotropic effect of immune/inflammatory pathways, an enhanced or attenuated inflammatory/immune response may substantially modulate the susceptibility to many metabolic, cardiovascular and neoplastic diseases (Bengoa, 2003, Diakos et al., 2014).

In summary, our functional analyses indicate HUI genomes contain highly divergent SNVs with a potential functional impact in genes participating in connected cellular processes. In many cases, the genes identified with VPFIs might explain a differential predisposition of Mapuche-Huilliches to metabolic, inflammatory or cancer-related diseases (Miquel et al., 1998, Carey et al., 2002, Nervi et al., 2006). It is important to mention that not all VPFIs found may represent increased risk factors. For example, women with Amerindian ancestry are at lower risk to develop breast cancer as compared to other ethnical groups (Fejerman et al., 2013, Fejerman et al., 2014). Moreover, several genes in the network have known functional relationships and have been associated with different types of diseases or disorders. This result suggests that accumulation of VPFIs in functionally related genes might modulate susceptibility of Native Americans to certain common diseases or disorders. Given that the HUI population along with other populations of Native Americans remain poorly investigated by genomic studies, the present study represent an important resource providing an ancestral population reference panel for future population-based studies on traits of interest (i.e. GWAS with a Native American SNV panel), as well as early diagnostic and prevention tools designed specifically for Native American populations.

**Genome-wide association study (GWAS) for Gallstone disease in the Chilean population**

Cholesterol gallstone disease is a frequent and economically highly relevant health problem in the world and one of the most prevalent gastrointestinal disorders in Chileans. Several world populations have been the subject of genetic analyses including genome-wide association studies (GWAS) resulting in a list of candidate genes associated with GSD (Lithiasis genes). However, genetic variation associated with these genes have only partially explained the genetics of GSD. The Chilean population has one of the highest prevalences of GSD in the world (36% and 17% in adult women and men, respectively) with a high genetic susceptibility. While Chilean genetic variation might represent an attractive source of novel susceptibility factors for GSD, this population has been the target of limited genetic studies. Here, we present results from the first GWAS performed on a Chilean cohort comprising 1,095 individuals (529 cases, 566 controls), aiming to uncover novel susceptibility risk factors for GSD in Chileans. All individuals were recruited according to their status on GSD (affected cases or healthy controls) diagnosed by abdominal ultrasonography. Genotyping was performed using the Affymetrix LAT 1 World array Plates. Genome-wide Imputation was performed with IMPUTE2 using the 1000 Genomes Project Phase 3 reference panel. Association analyses were done using SNPTEST. After quality controls and imputation, we analyzed ~9.5 Million SNVs observing a number of clear signals surpassing the suggestive genome-wide significance threshold ($<1x10^{-5}$). We detected a signal in chromosome 2 inside the ABCG5/8 locus ($p=7.04x10^{-6}$) which is a previously known GSD risk factor. Novel signals

were found on chromosome 7 ($p = 5.93 \times 10^{-7}$) and 14 ($p = 5.31 \times 10^{-6}$), in regions that encompass ELMO1 (gene involved in cell motility and previously associated to liver cirrhosis and diabetic nephropathy) and TRAF3 (involved in autoimmune response), respectively. Since women have an almost 3 fold increased risk for developing GSD compared to men, we performed a women specific analysis (489 cases and 525 controls). Results show a promising signal in chromosome 7 inside the GPR30 gene ($p=5.78 \times 10^{-6}$), a G-protein coupled estrogen receptor. Interestingly, previous deletion experiments of this gene in mice model have shown increased susceptibility to GSD.

Our results identify previously known as well as novel genes associated with GSD providing new insights into the genetic susceptibility of this prevalent disease. Replication and meta-analysis efforts with other independent cohorts are currently underway in order to have a more comprehensive view of the current results.

**Exome-capture sequencing strategy identifies a novel susceptibility locus for early-onset cholesterol gallstone disease in Chilean families.**

Besides its high prevalence, GSD in Chileans is characterized by its early onset age (Miquel et al.,1998). In order to uncover new genetic risk factors for early onset of GSD, we perfomed an exome-wide analysis of Chilean families with early onset GSD (children index cases). Twenty-four DNA samples (16 with and 8 without GSD) coming from 6 Chilean families including 6 young GSD index cases, their parents and second relatives were analyzed using the TruSeq Exome Enrichment kit followed by Illumina HiSeq2000 sequencing. Alignment to the reference genome (GRCh37) and variant calling were performed using SAMtools software. After annotation, filtering strategies and validation, a rare variant located within exon 1 of the ANGPTL4 gene (Angiopoietin like 4; rs186754194, chr19:8429328) was found in 2 independent families segregating significantly with the disease in one family ($p= 0.008$). This non-synonymous variation has a higher frequency in Latin Americans (MAF, 2%) and our cohort (20.8%) compared to rest of the world (0.46%) and the gene has been previously associated with lipid metabolism disorders. Our results show a novel associated gene with GSD providing new insights into the genetic susceptibility of this prevalent disease.

**Burden Analysis of Copy Number Variation in Gallstone Disease**

As mentioned above, specific genetic variants (SNVs) explain only part of the estimated heritability of GSD, suggesting the existence of additional genetic variants that remain to be discovered. In the present study we explored the contribution of copy number variants (CNVs) to GSD susceptibility. Two large case control cohorts composed of 1,076 Chilean and 1,957 German samples, in total 1,574 cases and 1,459 controls, were genotyped on the Affymetrix GeneTitan platform using the Axiom LAT 1 and GW Hu SNP World arrays, respectively. The Affymetrix Axiom CNV tool was used to calculate allele intensity ratios and B allele frequencies. This data was analyzed with the Nexus Copy Number Software to call CNVs using the SNP-FASST2 Segmentation algorithm. We considered CNV calls with at least 50 probes involved and larger than 100 kb. Further, we avoided calls overlapping telomeric and centromeric regions. A total of 1,414 events were detected (633 in Chileans;

781 in Germans), 86 of which (6.08%) were considered novel. Burden analyses of deletions and duplications between cases and controls revealed enrichment of samples with at least 1 duplication overlapping genes expressed in disease related tissues, particularly in small intestine that were suggestive for the Chilean cohort ($p=4.91x10^{-2}$) and statistically significant in the German dataset ($p=4.91x10^{-4}$). Analyzing the whole dataset together confirmed the finding ($p=2.10x10^{-4}$). Breaking the dataset down for gender effects revealed that the observation was almost exclusively male-specific. At this level the overall amount of CNVs detected was higher in cases than controls ($p=2.34x10^{-4}$) and primarily affected genes expressed in the small intestine ($7.96x10^{-5}$) with significant participation of genes involved in lipid metabolism according to gene ontology (GO:0006629, $p=1.6x10^{-3}$). Our preliminary results suggest CNVs may explain part of the missing heritability in GSD.

**B. Fruit crop genomes.**

We have been working on genomics of grapevine, peach and sweet cherry. Based on the experience acquired in the Genomic sequencing of `Sultanina´ we sequenced two table grape varieties used as parents in a breeding program. The aim of this work was to look for polymorphisms that could be used as molecular markers for traits of interest and then utilized for marker assisted selection. A great deal of polymorphisms was obtained and we are currently analyzing the data.

In a second project, the production and export of sweet cherry has increased exponentially in the last eight years in Chile. However, the production of this fruit crop is being heavily affected by the climate change. The blooming of the sweet cherry varieties cultivated in our country depends on an average of 1,000 chilling hours. In the last decade the winters in Chile are becoming milder with more frequent spring frosts that are affecting fruit production. The development of new varieties more adapted to Chilean conditions or molecular tools to assist breeding will positively impact the local fruit industry. Due to the lack of a publicly available sweet cherry genome, we sequenced the genome of a sweet cherry variety. The obtained assembly is an essential tool for the studies of the epigenetic regulation related to the dormancy and chilling requirement of flower buds in sweet cherry. The results obtained indicated that at least two loci coding for transcription factors that inhibit bud break suffer DNA methylation in response to prolonged cold exposure leading to their silencing and consequently dormancy release.

In a third effort, we have continued our work on the peach genome. Chile is the largest supplier of peaches from the Southern Hemisphere. Today, the main Chilean peach buyer is China and the fruit travel several weeks to arrive there, leading to postharvest disorders that affect fruit quality. In order to get a better understanding of the changes that take place during this process we performed transcriptomic analyses of three peach varieties exhibiting different harvest times and subjected to different postharvest treatments. These results allowed us to elucidate metabolic adjustments that are occurring in the fruit, affecting its final quality. The candidate genes identified can be used to develop molecular markers for assisted breeding.

## C. Fish Pathogens

### Transcriptome study of fish response to *Piscirickettsia salmonis* infection.

Piscirickettsiosis or Salmonid Rickettsial Septicaemia (SRS) is a bacterial disease that has a major economic impact on the Chilean salmon farming industry. Despite the fact that *Piscirickettsia salmonis* has been recognized as a major fish pathogen for over 20 years, the molecular strategies underlying the fish response to infection and the bacterial mechanisms of pathogenesis are poorly understood. We analysed and compared the head kidney transcriptional response of Atlantic salmon (*Salmo salar*) families with different levels of susceptibility to *P. salmonis* infection in order to reveal mechanisms that might confer infection resistance.

We ranked forty full-sibling Atlantic salmon families according to accumulated mortality after a challenge with *P. salmonis* and selected the families with the lowest and highest cumulative mortalities for microarray gene expression analysis. A comparison of the response to *P. salmonis* infection between low and high susceptibility groups identified biological processes presumably involved in natural resistance to the pathogen. In particular, expression changes of genes linked to cellular iron depletion, as well as low iron content and bacterial load in the head kidney of fish from low susceptibility families, suggest that iron-deprivation is an innate immunity defence mechanism against *P. salmonis*. To complement these results, we predicted a set of iron acquisition genes from the *P. salmonis* genome. Identification of putative Fur boxes and expression of the genes under iron-depleted conditions revealed that most of these genes form part of the Fur regulon of *P. salmonis*.

This study revealed, for the first time, differences in the transcriptional response to *P. salmonis* infection among Atlantic salmon families with varied levels of susceptibility to the infection. These differences correlated with changes in the abundance of transcripts encoding proteins directly and indirectly involved in the immune response; changes that highlighted the role of nutritional immunity through iron deprivation in host defence mechanisms against *P. salmonis*. Additionally, we found that *P. salmonis* has several mechanisms for iron acquisition, suggesting that this bacterium can obtain iron from different sources, including ferric iron through capturing endogenous and exogenous siderophores and ferrous iron. Our results contribute to determining the underlying resistance mechanisms of Atlantic salmon to P. salmonis infection and to identifying future treatment strategies.

### Complete genome sequence of *Piscirickettsia salmonis* LF-89

Only a limited amount of information is available on *P. salmonis* genome structure and mechanisms of pathogenesis. Nowadays, there are eight *P. salmonis* sequencing projects available at NCBI, four of them correspond to the LF-89 strain and four to environmental isolates; however they are all draft genomes with contig numbers between 227 and 534. The *P. salmonis* genome contains a high number of insertion sequences (N=179), which greatly enhances the complexity of genome assembly. By applying the strategy described below, we have now assembled the first complete genome sequence of this bacterium, providing a genetic background to understand *P. salmonis* biology and pathogenesis.

Genome sequencing of *P. salmonis* LF-89 was performed using one lane of an Illumina GAIIx instrument and two single-molecule-real-time (SMRT) cells of the PacBio RSII instrument. The resulting assembly consisted of four circular scaffolds; a complete chromosome of 3,184,851 bp and three plasmids named as pPSLF89-1 (180,124 bp), pPSLF89-2 (33,516 bp) and pPSLF89-3 (51,573bp).

Average GC content was 39.68 mol% for both the chromosome and the plasmids. We identified 74 RNA genes (18 rRNA and 56 tRNA) and 89 pseudo-genes. A total of 2,850 protein-coding genes were predicted, 2,634 were contained in the chromosome, 138 in plasmid pPSLF89-1, 35 in pPSLF89-2 and 43 in pPSLF89-3. Of the total predicted CDSs, 58.91 % had a functional prediction and 58.84 % were assigned to COG functional categories.

**D. Biomining Bacteria**

Copper is one of the most widely used metals in history. Copper is obtained mostly from copper sulfide ore processing, based on physical and chemical procedures limited to medium and high-grade ores. However, there are valuable resources of relatively low-grade minerals that, for economic reasons, cannot be processed by conventional methods and require other procedures for their exploitation. As a world leader in copper production, Chile has launched policies to improve the use of microorganisms capable of oxidizing iron and reducing inorganic sulfur compounds, which play an essential role in releasing copper from the mineral. As part of collaboration between the CRG, the Center of Mathematical Modeling (CMM) and the National Copper Corporation of Chile (BioSigma-CODELCO), during the last five years we have worked in study the molecular basis of the most efficient biomining bacterial consortium. Using complete sequencing and functional annotation, we have studied the metabolic capacities of the consortium, able to describe the potential of each biomining bacterium, some genome regulatory mechanisms and the metabolites required during the copper bioleach process. In addition, using next-generation sequencing we have released to date, five new different native Chilean bacteria genomes: Acidithiobacillus thiooxidans Licanantay, Leptospirillum ferriphilum Pañiwe, Acidithiobacillus ferrooxidans Wenelen, Sulfobacillus thermosulfidooxidans Cutipay and Acidiphilium multivorum Yenapatur.

All these works has been presented in international conferences, has produced 5 patents and 8 articles now published in well known biotechnology impact factor journals. In terms of the international collaboration, jointly with the “Institut National de Recherche en Informatique et en Automatique” INRIA-France, we are developing a new mathematical model to reconstruct global metabolic and gene regulatory networks in biomining organisms. This collaboration provided an active participation of members of the INRIA, who visited Chile in several times during the last five years.

**E. The genome of *Salmo salar***

Our center has participated in the executive scientific committee in the consortium that sequenced the Atlantic salmon, *Salmo salar*. More precisely, we have worked on different tasks concerning the quality control of the process of data release and sequence assembly as

well as producing the AllPath assembly which was used to consolidate the assembly produced using Masurca assembler in JCVI and U. Maryland. The last assembly of the salmon sequence which also identifies and separates chromosomes was released during the first semester of 2015. In parallel, we have developed our own annotation of the genome that is publicly available through our database SalmonDB2.0. This database was recently included in a community article published in Nucleic Acids Research to highlight the advantages of the Biomart platform. Finally, we have assisted Aquainnovo, a Chilean biotech startup company, to develop a novel high-density SNP panel which provides an excellent platform for the dissection of economically relevant traits, assisting breeding programs through genomic selection and genetic studies in both wild and farmed populations using high-resolution genome-wide information from Atlantic salmon.This array was used to genotype 480 fish representing wild and farmed fish from Europe, North America and Chile. An article entitled "High-throughput (SNP) discovery in Salmo salar): validation in farmed and wild American and European populations" was recently submitted.

**Research Line 3. Gene Expression in Cells**

**a. Epigenetic mechanisms**

*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 (PcG) and Trithorax/COMPASS (Trx) complexes. These complexes are recruited to target genes where they can either favor silencing or activation (respectively) of transcription during differentiation. In collaboration with national and foreign groups, we have analyzed the contribution of the PcG and Trx proteins during both hippocampal development and the mesenchymal-osteogenic transition exploring their specific role during transcriptional control of genes that are critical for establishing and maintaining both phenotypes (Bustos et al. 2013; Rojas et al. Submitted). We found that the activity of the PcG and Trx complexes at target genes occurs concomitantly with the presence of a specific pattern of epigenetic marks at histones as well as with binding of additional epigenetic regulators to promoter sequences that surround the transcriptional start site (see diagrams in figure 4, annex). We are currently exploring the molecular mechanisms, including signaling pathways, that support the role of these regulatory complexes either during activation or repression of these phenotypic genes.

To further understand the mechanisms underlying PcG- and Trx-mediated control of mammalian osteogenic lineage commitment, Dr. Gino Nardocci (post-doctoral fellow) has also examined the expression of Long-Non-Conding RNAs (LncRNAs) during osteoblast differentiation, focusing on the identification of new LncRNAs that are associated with either PcG and Trx complexes. Our search has led to the identification of 31 new LncRNAs (see figure 5, annex) which are currently evaluated through gain- and loss-of-function experiments to assess their role during PcG- and Trx-mediated control of osteoblast-specific gene transcription.

In collaboration with Dr. Gutierrez (CGR) and Dr. van Zundert (UNAB) we have been also defining the role of specific miRNAs (miRs) in controlling gene expression during

hippocampal neuronal differentiation. Dr. Laura Guajardo (post-doctoral fellow) has used microarray and deep-sequencing analyses, to define a reduced population of miRs that are differentially expressed during maturation of hippocampal neurons and that are predicted to interact with sequences at the 3'-end of mRNAs expressed during this neuronal maturation process. Using bioinformatics strategies, we have elaborated interacting networks between these selected miRs and their potential mRNA targets in hippocampal neurons (see figure 6, annex), which have led to the identification of new potentially relevant functional interaction nodes. The experimental verification of these nodes is currently under analysis by gain- and loss-of-function studies.

**b. Stem cells**

*Mesenchymal stem cells (MSCs) as therapeutic agents for tissue engineering applications*

It has been shown that in pathological conditions including cancer, birth defects and trauma due to accidents, there may be a massive loss of bone tissue. Therefore, there is a necessity to develop new approaches for bone tissue engineering. Collaborative work between Dr. Montecino and Dr. Allende (Rojas et al. submitted) demonstrates that the JARID1B histone demethylase functions as an epigenetic brake for osteogenic-lineage commitment of mesenchymal cell lines, due to its role as a mediator of transcriptional silencing of the osteoblast master gene Runx2. Human mesenchymal stem cells from the umbilical cord (e.g. Wharton Jelly-derived cells, WJ-MSCs) exhibit high therapeutic potential in humans given their multipotency, angiogenic capacity and immunomodulatory properties. Hence, work performed in collaboration with Dr. Palma (CRG) has been evaluating whether a loss-of-function of JARID1B can favor the osteogenic conversion of uncommitted WJ-MSCs. Our results (not shown) show that in WJ-MSCs JARID1B is expressed constitutively. As a proof of concept we have also shown that JARID1B expression may be significantly decreased via infection with lentiviral particles carrying a shRNA against JARID1B in the osteosarcoma cell line SAOS and that this JARID1B knockdown results in a significant increase in Runx2 expression. We are currently carrying transcriptomic analyses to further confirm similar results as well as a change in terms of osteogenic commitment reached by WJ-MSCs.

**c. Research on Niemann -Pick diseases**

Niemann-Pick (NP) diseases are hereditary lysosomal diseases caused by mutations in the *NPC1* or *NPC2* genes (NPC) or the gene coding for the Acid Sphingomyelinase (ASM) enzyme (NPA and NPB).

Niemann-Pick type B (NPB) disease is one of the most frequent lysosomal storage disorders in Chile, with at least 45 confirmed cases. The A359D mutation has been described only in Chilean NPB patients. In collaboration with J.F. Miquel we found that the frequency of the A359D variant in the healthy Chilean population was be 1/105.7, predicting a disease incidence of 1/44,960 in Chile, higher than the incidence estimated by the number of confirmed NPB cases (Acuña et al., 2015). In collaboration with M. González and R. Gutierrez we determined the haplotype background of homozygous patients and found a conserved haplotype and a shared 280 Kb region around the ASM gene in 6 patients

analyzed, indicating that the variant originated from a common ancestor. The haplotype frequency and mitochondrial DNA analysis suggest an Amerindian origin for the variant and a common founder (Acuña et al., 2015).

Information on common variants such as A359D is essential to guide the successful implementation for future therapies and benefit to patients. This research line is part of the PhD thesis of Mariana Acuña, PhD student of the Medical Sciences Program from P. Universidad Católica de Chile and the results obtained were recently published (Acuña et al. 2015).

In collaboration with M. González we studied gene expression alterations in livers and cerebella from the NPC mouse model (Vázquez et al. Plos One 2011). In addition, we analyzed copper (Cu) transport and metabolism because it could contribute to oxidative stress in NPC nervous and hepatic tissues. We found increased Cu content in the plasma and decreased Cu levels in the bile of NPC mice, whereas we did not observe a significant change in copper content in the cerebellum (Vázquez et al. Biometals 2012). These results suggest a cell-type dependence of Cu accumulation in NPC and suggest that Cu transport imbalance may be relevant to the liver pathology observed in NPC disease. Accordingly, we found increased Cu content in a hepatic in culture NPC model, whereas no changes were detected in a neuronal NPC model (Vázquez et al. Biometals 2012). In addition, we analyzed the systemic consequences of Cu transport alterations in the NPC mice model analysing the consequences of low-Cu and high-Cu diets and Cu chelating agents on Cu homeostasis and cardiac alterations. This research line was part of the PhD thesis of Graciela Argüello, student of the PhD Nutrition Program from Universidad de Chile and the results were published last year (Argüello et al. 2014).

**d. Role of the cystein-serine rich nuclear proteins (CSRNP) in transcription regulation and genome expression.**

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 obtained RNA-seq data from loss and gain of function condition of DAxud1 and we will combine it with from a DamID experiment (using Axud1-Dam fusion) in D. melanogaster to identify its target genes and to gain insights about the function of these proteins. We also performed a Mass spec determination of the proteins immunoprecipitated by DAxud1. This biochemical information will serve to further understand the mechanism controlling transcription in the genes regulated by DAxud1.

**e. Selective translation in Drosophila cells.**

Translation starts at AUG codons, its selection relay on several factor among them assembly of initiator complex. The KEOPS complex is associated with a tRNA modification that ensures correct codon recognition and proper translation. We have performed in vivo disruption of this complex showing that it induces UPR, affecting also TOR activation and cell growth (Ibar et al., 2013; Rojas-Benítez et al., 2013; Rojas-Benítez et al., 2015). By ribosome footprinting we are investigating if variations in the levels of this modification affect AUG

selection and therefore the cellular translatome, thus emerging as an additional layer of control of genome expression.

**f. Regeneration of animal tissues.**

We have continued to generate data on the regenerative capacity in different animal models. In zebrafish, we have shown that axons in both the peripheral and central nervous systems, can effectively regenerate. The regrowth of nerve fibers to form functional nerves depends highly on the immune system and on glial cells that accompany neurons (Ceci et al., 2014). Furthermore, we have begun studies to show that some of the pitfalls of mammalian tissue regeneration can be overcome by the local production of oxygen in tissues, one of the main factors limiting recovery of organs in humans. We have engineered animal-plant chimeras by introducing genetically modified algae (*Chlamydomonas reinhardtii*) into both fish and mice. This work, recently published (Alvarez et al., 2015; Schenk et al., 2015), shows that it is possible to design a tissue engineering strategy that could be used in plastic surgery and tissue reconstruction in a clinical setting.

**g. Biological Networks**

We have worked in different problems concerning the genome scale reconstruction and analysis of regulatory and metabolic networks from the perspective of integrative biology. In addition, we have applied and validated our methods in new (non-model) organisms studied in our center.

*a.* Reconstruction of metabolic networks:

- We have conceived the method Pantograph to reconstruct the metabolism of an organism from genome data and a fine use of phylogenetic relations. With this method we wanted to be able to treat the non-model organisms that appear in our research of extreme regions in Chile. This method was validated in yeast but is already used to reconstruct the metabolism of biomining bacteria and the alga Nannochloropsis salina. Together with INRIA-Rennes we are improving Pantograph combining it with the protocol MENECO developed by them. The new ideas, that will produce MENECO++, integrate phylogenetic analysis with the modeling of the metabolism. Here we are trying to provide ideas in relation to the way an organism configure a network; this is stated as an optimization problem, usually very hard to solve so a complexity analysis is required.
- A second idea in this project was to integrate metabolic reactions with the way enzymes catalyzing them are located in the genome in order to define “functional units” controlling the metabolism. More precisely, we define the notion of shortest genome segments (SGS), which encapsulate regions of the genome that as a dense block catalyze the reactions present in the organism metabolism. In bacteria, the SGS are associated to more than 40% of the metabolism. Beyond many interesting features of the new concept that has been proved in E. coli we used this new notion in biomining consortium of bacteria proving that SGS of organisms are complementary to fill complete pathways interesting for the biomining process, so it enhances the particular features of the metabolism of an organism.

- Finally we have intensively worked in the study of metabolic precursors, in particular to list precursors sets. That is, given the metabolic network of an organism and a specific metabolic lens (objective reaction or metabolite), a precursor set is defined as a minimal set of external metabolites allowing the organism producing a metabolic defined target. During this period we have developed a new algorithm to enumerate all precursors set for a given objective. More precisely, this new method can automatically discard solutions that produce cycles that are unfeasible stoichiometrically. So this algorithm is an accurate method to obtain all feasible solutions. We evaluate and enhance the method by applying it to problems of center interest, such as understanding the relationship to metabolic rate that occurs between symbiotic bacteria Pitscirickettsia salmonis and their host the Atlantic salmon.

*b.* Reconstruction of transcriptional regulatory networks: The discovery and characterization of transcriptional regulatory elements within a network controlling changes in gene expression in response to environmental stresses is a central question in Systems Biology. Gene co-expression, evidenced by different correlation measures between genes in high throughput expression data, shows that transcriptional activity is coordinated and pinpoints to the role of regulation in the acclimation processes. State-of-the-art bioinformatics methods aim to locate putative transcription factors and their binding sites in the genomic sequence leading to putative transcriptional regulatory networks aiming to provide mechanistic hypothetical explanations for most of the observed co-expressions. These putative networks can be huge and unreliable since prediction models have low specificity. To solve these issues we proposed LOMBARDE, a method that integrates putative transcriptional regulatory networks with co-expression data to determine the simplest and most confident regulatory network explaining observed co-expressions in prokaryotes. In a test case using E. coli expression data of more than 900 microarrays, LOMBARDE produced an explanatory regulatory network coherent with most of the observed co-expressions that uses only 19% of the regulations in the initial putative network. Interestingly, LOMBARDE is biased to output regulations that have been experimentally validated, preserving most of them. The topological characteristics of the final network produced by LOMBARDE, such as average degree and centrality role of the main regulators, are close to the values described in the literature for networks of this type. In a second test application, LOMBARDE was used to extend the set of known E. coli regulations with a small number of additional ones, proposing a mechanistic explanation for co-expression data. Finally, we show that LOMBARDE results are robust to the choice of co-expression index and have most of the characteristics of a realistic transcriptional regulatory network.

*c.* Dynamic models of hybrid systems: during this project progress has been made to consolidate and improve our implementation of the simulation of biological processes for hybrid systems. We have focused on two aspects: the method and software.

Regarding the method, we have formally established the acclimatization process as defined scope and lines of work in bioleaching and cellular decisions.

**3.2.-Formation of human resources in the discipline**

Without hesitation, we can say that human resource formation is by far the most outstanding feature of the CGR when considering the output of the Centers of Excellence program as a whole. A combination of successful research projects, exciting topics and attractive interdisciplinary atmosphere are likely contributors to this outcome. Just considering the per capita numbers of students produced we can feel satisfied that our performance is what would be desired and expected from this scientific program. A total of 70 postdocs belonged to the CGR in the 2011-2015 period.

As the generation of a critical mass of young scientists working in the field encompassed by the CGR was one of our strategic aims, we have estimated our impact in this area by examining our performance on human capital insertion in academia. We can enumerate the following cases of CGR postdocs who have moved on to permanent positions:

1. Fernan Federici, Prof. at U Católica
2. Rodrigo Assar, Prof. at U de Chile
3. Christian Hodar, Prof. at U de Chile (incorporated as CGR Associate Investigator)
4. Vicente Acuña, Prof. at U de Chile
5. Igor Pacheco, Prof. at U de Chile
6. Rodrigo Pulgar, Prof. at U de Chile
7. Javier Canals, Prof. at U Austral de Chile
8. Dianna Gras, Prof. at CONICET, Argentina
9. Julian Verdonk, Prof. at Wageningen Univ, Holland

In other words, we have bred 9 new scientists of which 7 are now group leaders in Chilean universities. Of course, this number is expected to grow as more graduates leave the CGR.

Furthermore, we have recruited three young Chilean scientists that had achieved semi-independent status abroad and they are now working within the CGR on their projects:

1. Tomás Egaña, had a position at the Technical University of Munich, Germany. He was incorporated as an Associate Investigator at the CGR in 2012. Has just obtained a position at U. Católica and has his own funding.
2. Ricardo Nilo, was at Stanford University and is now leading two CGR projects which he will incorporate into his funding proposals and will apply for an independent position.
3. Gonzalo Olivares, also from Stanford U., he has an NIH RO21 grant which he has brought with him to the CGR from where he will look for a permanent position at a local university.

These individuals represent our pioneering effort to reverse the "brain drain" that has been taking place in Chilean science, as hundreds of young people have been awarded fellowships to get PhDs and postdocs abroad without a clear mechanism for their return. We hope to be able to contribute further to this highly urgent task. We are including these types of metrics in our proposal for renewal as a way to measure our performance and impact.

One of the requests of the FONDAP evaluating committee was to stimulate interactions among the postdoctoral fellows at the Center. Ideally, they should organize meetings independently of the PIs and seek out ways in which to collaborate that may escape the general research planning at the CGR. A first meeting of the postdocs in 2015 gathered 45 of them in an event that took place during an entire day. There were scientific presentations (oral and posters) and invited speakers from government and student organizations. A report describing the results of the meeting and discussion that took place there was submitted to the PIs. A summary of this report is included in the Annex.

The number of PhD and Master's theses carried out with CGR investigators also exceeded our expectations. 118 graduate students were trained in our Center, a remarkable number of young scientists that represent one of our strongest assets.

Our Center has applied for funds to establish a Research Training Group together with the PhD program from the Center for Organismal Studies at the University of Heidelberg. This agreement and funding scheme will allow us to link the PhD program at UH with the PhD Program in Cell and Molecular Biology at the University of Chile. The funds include exchanges for students, visiting professors and workshops to be implemented bidirectionally.

The work performed under the umbrella of the CRG at Andres Bello University has strongly supported the research activities of two doctoral programs associated with this institution. Thus, students from the Doctoral Program in Molecular Biosciences and the Doctoral Program in Biotechnology have carried out both research units and thesis work in the laboratories associated with CRG. This support was an important component during the successful process of national re-accreditation that both programs faced during 2013 and 2014. Additionally, students from other institutions, not associated with CRG, including University of Concepcion and Austral University of Chile, have also performed research units and doctoral theses in these same labs with support of the CRG. Noteworthy is that in 2014, a student from the Doctoral Program in Biological Sciences of the University of Concepcion (Fernando Bustos), who performed his thesis under the joint direction of Dr. Brigitte van Zundert (UNAB) and Dr. Martin Montecino (UNAB-CRG) was awarded with the price to "The Best Doctoral Thesis of the Year in Biological Sciences". This price is awarded yearly by the Chilean Society for Cell Biology together with the Chilean Foundation for Cell Biology.

Two graduate students performed their PhD thesis in research lines related with Niemann-Pick diseases. One of them was co-advised between M. González and S. Zanlungo. The second one, still under development, related with Niemann-Pick B disease,

has been positively influenced by the Center expertises such as bioinformatic analysis and 3D modeling (provided by R. Gutierrez, M. Gonzalez and M. Latorre).

### 3.3.-National and international collaboration

We have sought to become visible to the international scientific community through our wide range of collaborators and partners all over the world. We continue to publish extensively with foreign colleagues and several of our projects are carried out in partnership with colleagues from centers in North America, Europe and Asia/Oceania. Our investigators are consistently and frequently invited to speak at the major international forums and congresses. We highlight some of the activities which can be considered a reflection of the status we have achieved outside our borders.

We are establishing formal (and funded) collaboration agreements with several institutions. We have a signed agreement with the Universidad de La República in Uruguay, which, together with the Institut Pasteur of Montevideo, are partners in the Austrolebias genome project. We believe it is the first genome project undertaken by two South American countries. We also have an agreement with the ARMI, Australian Regenerative Medicine Institute, of Monash University at Melbourne. In addition to exchanges, we held a symposium at the CGR with ARMI scientists; recently, Dr. Allende, the CGR Director, was named Adjunct Professor at Monash University. We have strong ties with the University of Heidelberg, specifically with the Center for Organismal Studies, together with whom we organized a very successful international course in biological imaging technologies. We have applied for funding to establish a joint graduate program which we expect to implement starting 2017.

At the CGR, we have strong ties with France, both institutionally and individually. Our bioinformatics group has had a strong collaboration for more than five years with INRIA-France, in particular with the teams of Bamboo in Lyon and Dyliss in Rennes. We are an associate team of INRIA-France through the project IntegrativeBioChile. The selections of the associated teams is done considering the scientific goals of each project as well as the exchange program and the quality of partners. In particular, IntegrativeBioChile has used mathematical and computational methods in order to explore and integrate heterogeneous biological data and be able to produce reliable interaction networks at regulatory and metabolic level. This project has been extended to Chile via INRIA-Chile where our Center is the leader of the "Omics Sciences" research line. It includes genomics, proteomics, transcriptomics and metabolomics. Our aim is to develop a bioinformatics platform and to structure a set of consulting services based on modules to integrate and analyze large sets of heterogeneous omic data, in order to produce networks of biological interaction and biomarkers involved in a productive system.

With support of CRG the laboratory of Dr. Martin Montecino has been able to establish and maintain tight collaborations with research groups located within and outside of the country. Among them is the work carried out with the group of Dr. Brigitte van Zundert at UNAB and the team lead by Dr. Marianne Rots at University of Groningen, The Netherlands. This collaboration allowed student exchanges between both sides and the generation of new

artificial transcription factors that modulate epigenetic mechanisms and transcription of genes associated with neuronal plasticity (manuscript in preparation). Also important has been the collaboration with Dr. Andre van Wijnen at Mayo Clinic, USA, to address epigenetic mechanisms that control gene expression during osteogenic differentiation. Several manuscripts already published or in revision, demonstrate the success of this scientific interaction. Additionally, Dr. Montecino's group has teamed with Drs. Gary Stein, Janet Stein, and Jane Lian, at University of Vermont Medical College, to address mechanisms that control gene transcription in osteoblastic cells. This collaborative effort has generated a significant number of joint publications as well as an important number of scientific exchanges between both parties. It is also important to note that collaborations with groups located within the country have led to several joint publications. These groups include those lead by Dr. Manuel Kukuljan (University of Chile), Dr. Juan Olate (University of Concepcion), Dr. Mario Galindo (University of Chile), Dr. Katherine Marcelain (University of Chile), Dr. Marcela Hermoso (University of Chile), Dr. Francisco Aboitiz (P. Catholic University of Chile) and Dr. Christian Gonzalez-Billault (University of Chile).

Three international collaborations have been strengthened in recent years with researchers working in lysosomal diseases; Tony Futerman (Weizmann Institute, Israel), Frances Platt (University of Oxford, UK) and Edward Schuchman (Mount Sinai School of Medicine, NY, USA). With the last two researchers we have published manuscripts in collaboration with researchers from the Center (Argüello et al. 2014, Acuña et al. 2015)

### 3.4.- Outreach:

We summarize all of our Outreach and Communications activities in a separate report included in the Annexes.

**IV. OTHER RELEVANT ASPECTS:** Analyze the effects that the creation of the Center produced on:

Changes in the research lines.

This was discussed in previous reports. After the first year and comments from the review panel, we made significant changes to the scientific objectives of the CGR. The new objectives and research lines have stayed in place since then without modification. In terms of publications, most lie within the scope of objective 3, as discussed above. During the final stage of the five year period ending in December 2015, we expect several publications from objectives 1 and 2 to be forthcoming.

Other relevant achievements.

The CGR has been involved in the generation of two new environmental regulations that have just become public as of this writing. Our work with the zebrafish and its application as an efficient environmental monitoring tool has led us to propose the generation of these regulations to the Chilean authorities. One of the regulations was adopted from the current European standard for water quality assurance (an ISO standard) and the other is of our own creation. The significance of this achievement is that we have put in place standardized national tests that can evaluate the condition of continental waters in the country. Importantly, we have set up a specialized lab and fish facility (inaugurated in June of 2015) that can carry out these tests for either private or public clients.

Infrastructure.

The CGR is a Center with no physical building to house it; we rent space for our offices and bioinformatics group but none of the labs are in close proximity to one another. This is obviously a drawback for collaborative work but we have overcome it in many ways. To strengthen the infrastructure at our labs, we have built during the past five years several facilities for plants, animals, computers and sequencing equipment. Funding for these improvements has come both from the CGR directly as well as from matching funds provided by the universities.

## V. CENTER PROJECTIONS

---

The CGR aspires to obtain five more years of funding to continue its mission and accomplish its objectives. Clearly, our results to date are significant but incomplete, considering what we yet have to achieve. The most important publications of the first half of the project are about to be generated. Maturation of the research lines and their transformation into relevant findings that can be recognized worldwide are in the near to mid future. Most importantly, the young investigators that have been trained at the CGR look to our center as a place of opportunities and professional growth. In a separate document (the continuity proposal) we are presenting the plan we have envisioned for continuing the work we are now embarked on and introducing new challenges, that have become evident at this point in time.

As a summary of the achievements in terms of numbers, we present a final table of indicators, one we have provided in our yearly reports. **Please note that the actual values for 2011-2015 only include half of 2015, underestimating our productivity for the five year total.** The expected values on the right hand column were declared in our original proposal in 2010.

| Indicator | 2011-2015 | Expected for 5 years |
|---|---|---|
| **Number of ISI papers** | 155 | 179 |
| **Total Impact Factor of ISI papers** | 763.833 | 800 |
| **Average Impact factor of ISI papers** | 4.93 | 4.0 |
| **5 year citations (all papers)** | 5267 | 2800 |
| **5 year citations (papers from 2010-2015)** | 1666 | 1608 |
| **Co-authored publications** | 28 (18%) | 20 |
| **Postdocs associated to CGR** | 70 | 50 |
| **PhD students associated to CGR** | 73 | 73 |
| **Total number of theses directed** | 136 | n/a |
| **Co-directed theses** | 8 | 12 |
| **Congress presentations** | 582 | n/a |
| **Conferences and courses organized** | 43 | n/a |
| **Postodcs with permanent positions obtained** | 10 | n/a |

**VI. FULFILLMENT OF THE INSTITUTIONAL COMMITMENT:**

The institutions harboring the CGR have complied with the commitments that they agreed upon at the start of the project. The in-kind contributions they have made include salaries for personnel and postdocs, travel funds and infrastructure.

**VII. ADVISORY COMMITTEE:** Indicate the way in which the committee contributed to the development of the Center. Discuss the most relevant problems found in carrying out this endeavor.

The advisory Committe has met with us once a year to evaluate our performance and make suggestions for improvement. The Committees' valuable help has allowed us to focus on the most relevant projects (Center Projects) facilitating the organization of tasks and responsibilities. The members of our Committee are:
Prof. Enrique Lessa. Universidad de la República, Uruguay
Prof. Ben Koop. University of Victoria, British Columbia, Canada
Prof. Alan Benett. University of California, Davis. USA.

**VIII. TABLES:** Using the attached form, list all Center publications, congress and seminar presentations, courses, materials and other activities of dissemination during the first five -year period of the Center 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 that has not been sent to CONICYT in past reports.

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 that has not been sent to CONICYT in past reports.

3. **Books and book chapters**

- ✓ Include a hard copy of every **PUBLISHED** book that has not been sent to CONICYT in past reports.
- ✓ Include a digital copy of the front page of the chapter in the case of a book chapter that has not been sent to CONICYT in past reports.

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 that has not been sent to CONICYT in past reports.

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 that has not been sent to CONICYT in past reports.

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.