Final report for project: **Globally connecting single nucleotide variations to transcriptional regulation in vivo ("SNPs and TF binding")**
Period covered: 1/10/2009 – 30/09/2012

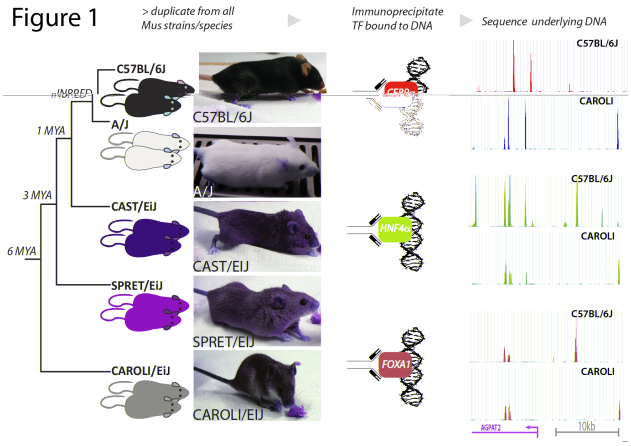
Summary of proposed objectives:

The overarching goal of this project was to understand how subtle changes in regulatory genetic sequence could generate phenotypic differences between closely related species via influencing binding of regulatory proteins (transcription factors, TFs) to the DNA. Changes in TF-DNA interaction have, in turn, impact on the recruitment of complexes that transcribe the relevant genetic information. My aim was to interrogate the drivers responsible for moderating the impact of the sequence changes on TF binding. Understanding how sequence divergence in non-coding parts of the genome relate to the phenotype differences in closely related species can help us understand the origin and driving forces behind phenotypic diversity.

Aim 1: *Identify the in vivo binding sites of regulatory proteins (transcription factors, TFs) within the genomes of closely related inbred mouse strains.*
Aim 2: *Connect the changes in TF binding location to the changes in the underlying and surrounding sequence. Propose the possible causative link between the sequence and TF binding changes for each of the locations genome-wide.*
Aim 3: *Test our hypothesis by eliminating one of the drivers of TF binding change.*

Summary of achieved objectives:

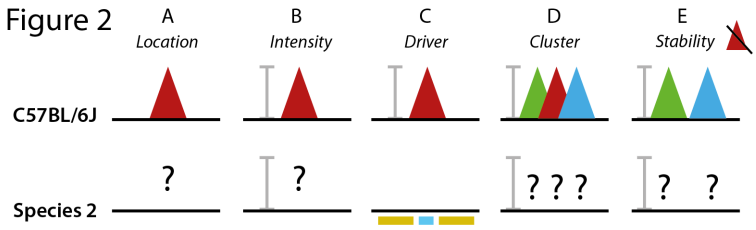
Aim 1: I have experimentally determined more than 50,000 binding sites genome-wide for three transcription factors (TFs: CEBPα, HNF4α, and FoxA1) in mouse species separated by less than 6 million years of evolution, all in duplicate, to acquire the highest quality data, allowing me to achieve an unprecedented resolution for interrogating TF binding among closely related mammals (**Figure 1**). By high-throughput sequencing of the DNA fragments each TF interacts with, I was able to extract two crucial parameters: both the precise and reproducible *in vivo* binding locations for each TF, as well as the intensity of the binding, representing the average probability that the TF binds to that genomic location within the population of liver cells for each species.



Aim 2: First, in order to assess the initial steps in the evolution of transcriptional regulation, we compared how well both the location and binding intensity for each *in vivo* TF binding site is preserved over 6MY of evolution, representing <3% of sequence divergence. While a conserved location indicates that the same physical position between divergent mouse species is continuously occupied (**Figure 2A**), similar intensity represents the probability that the transcription factor binds to this location in a similar number of liver cells (**Figure 2B**). We observed an extensive change in both the physical location and the binding intensity among the five closely related mouse species for all three transcription factors. While changes in the short

sequence that is directly bound by each transcription factor (**Figure 2C**) can explain a significant fraction, it cannot explain more than a third of all the changes in TF binding we observe between the mouse species. In addition, our experimental design allowed us to interrogate binding of all three transcription factors simultaneously (**Figure 2D**). We found that TFs that bind to the same location exhibit coordinated changes in binding intensity. This implies that the driver behind changes in TF binding acts locally and often affects the whole cluster of TFs.

Aim 3: Since the intensity of TFs bound within the same cluster undergoes coordinate changes during evolution, this suggests that the binding of one TF has the potential to influence binding of the other TFs within the same cluster. I have tested the hypothesis by using a mouse that has one of the TFs (CEBPα) excised from the genome (a knock-out mouse). While CEBPα is an important TF for liver development and function, we do not see a profound effect on the stability of clusters of TF binding containing all three TFs (**Figure 2E**). These clusters are bound in most of the cells and are evolutionary stable, containing possibly many more TFs that may be better candidates for the pioneering function of chromatin opening and recruitment of TFs other than CEBPα. Possibly, other TFs, similar to CEBPα (CEBPβ and CEBPγ) can compensate for CEBPα loss. I am currently testing if excising HNF4α will have similar or drastically different effect.



Dissemination activities and related projects:

Meetings and awards: In the last 3 years, I have attended six meetings (all highly relevant to understanding the transcriptional regulation field) and presented my work as a poster at two meetings (the EMBO meeting in Vienna 2011 and the Systems Biology meeting in Cold Spring Harbor in 2012). Also, my collaborator, David Thybert, obtained "The Best Poster" prize for a poster that was solely based on the work described above, presented at the Biology of Genomes 2012 conference in Cold Spring Harbor. In addition to that, an extension of this grant secured me a non-stipendiary fellowship at Hughes Hall College as well as a one-year fellowship from CRUK that is currently securing the funds necessary for bridging my current Postdoctoral position and a permanent position.

Publications: The work outlined in the proposal is currently ready for submission to the journal Cell and I share my first authorship with my computational collaborator David Thybert. I have been involved in number of collaborative projects, all relevant to the outlined proposal. I am intimately involved in the sequencing of the Caroli genome as well as obtaining data aiming to capture the functional consequences of changes in TF binding. These efforts are currently being finalized for publication on which I will be the co-first author. Additionally, in collaboration with two PhD students, Sarah Leigh-Brown and Angela Goncalves, I have explored what are the main drivers behind regulating transcription and splicing in mouse sub-species. Also, I contributed experimental design and data to the project led by Stephen Watt and Claudia Kutter, exploring the regulation of long non-coding RNA in rodents.