Project Title: Design of synthetic plant and mammal gene regulatory networks using nonparametric Bayesian approaches

1. Summary

Despite some impressive success stories, synthetic biology approaches still require a significant element of trial and error. Our ability to design and predict is still poor. This is largely due to the challenge of simulating large networks and using them effectively to inform experiments. Here, we aim to develop a suite of computational tools for the rational data-driven (re)design of gene regulatory networks (GRNs) that can be used to guide synthetic biology efforts. Our approaches will be validated *in silico,* and using data from well characterised plant and mammalian systems. As a proof of principle, GRNs associated with stress responses in *Arabidopsis thaliana* will be synthetically rewired and tested in protoplasts. In parallel, we will develop approaches to identify key regulators that will allow us to rationally direct cell fate decisions in human embryonic stem cells.

2. Report and outcomes

In this report, we investigated two approaches to the design of synthetic interventions in either plants or mammals.

In section 2.1. we investigate the design of gene regulatory networks using Bayesian nonparametric approaches to emulate dynamical systems. We first investigate the use of these methods on *in silico* datasets and subsequently use these approaches to design network rewiring in plant systems.

In section 2.2. we address approaches for rewiring networks in mammalian systems. Here we make use of single cell RNA-seq datasets measured over several developmental stages. Provided that these single cell observations can be ordered over a continuous trajectory, they represent a high-resolution (pseudo) time-series. As part of this project we developed a Gaussian process model of branching, that allowed cells to be ordered over a continuous developmental program with multiple cell fates. Using this, we thus reverse engineer the dynamical system as a function of pseudotime, and identify a number of early regulators likely to be involved in the specification of primordial germ cells, embryonic precursors of the sperm and eggs. We test the impact of these regulators using small molecule inhibitors. All four of the regulators tested had a phenotype in an *in vitro* model that derived primordial germ cell-like cells from embryonic stem cells.

2.1. Dynamical systems inference

The plant response to biotic and abiotic stresses is characterised by extensive transcriptional reprogramming. This is a coordinated, systemic response involving changes in gene expression and regulation. Changes in the regulatory regions of genes have been crucial in plant domestication and diversification (Meyer et al., 2013; Century et al., 2008; Doebley et al., 2006). As such, rewiring - a change in the edges of a GRN - either through the introduction of new edges or the deletion of existing edges, affecting the regulation of the node - is an important mechanism for responding to the changes in the environment and developing phenotypes. We use rewiring to optimise the gene expression levels in a GRN in order to produce a certain phenotype. Before rewiring to optimise large plant networks, we first look at optimising smaller *in silico* networks.

2.1.1. Rewiring genes *in silico*

We use Gaussian process regression (GPR) (Rasmussen, 2004) in order to predict the effect of rewiring on gene expression in a network. We validate our method using DREAM4 challenge *in silico* networks (Marbach et al., 2009). These are time series expression datasets from 10-gene ODE networks.

Measurements over 21 time points are available for each of the 10 genes under a specific perturbation. The perturbation is applied to a few genes in the network for the first half of the time series and is removed for the second half of the time series. As these ODE functions are known, they can be used to generate “true” data for rewired networks that our GPR predictions can be compared to.

GPR learns the underlying function connecting the expression of regulators to that of the targets at preceding time points, uses these functions to predict the levels of genes under the same perturbation but with a different network conformation.

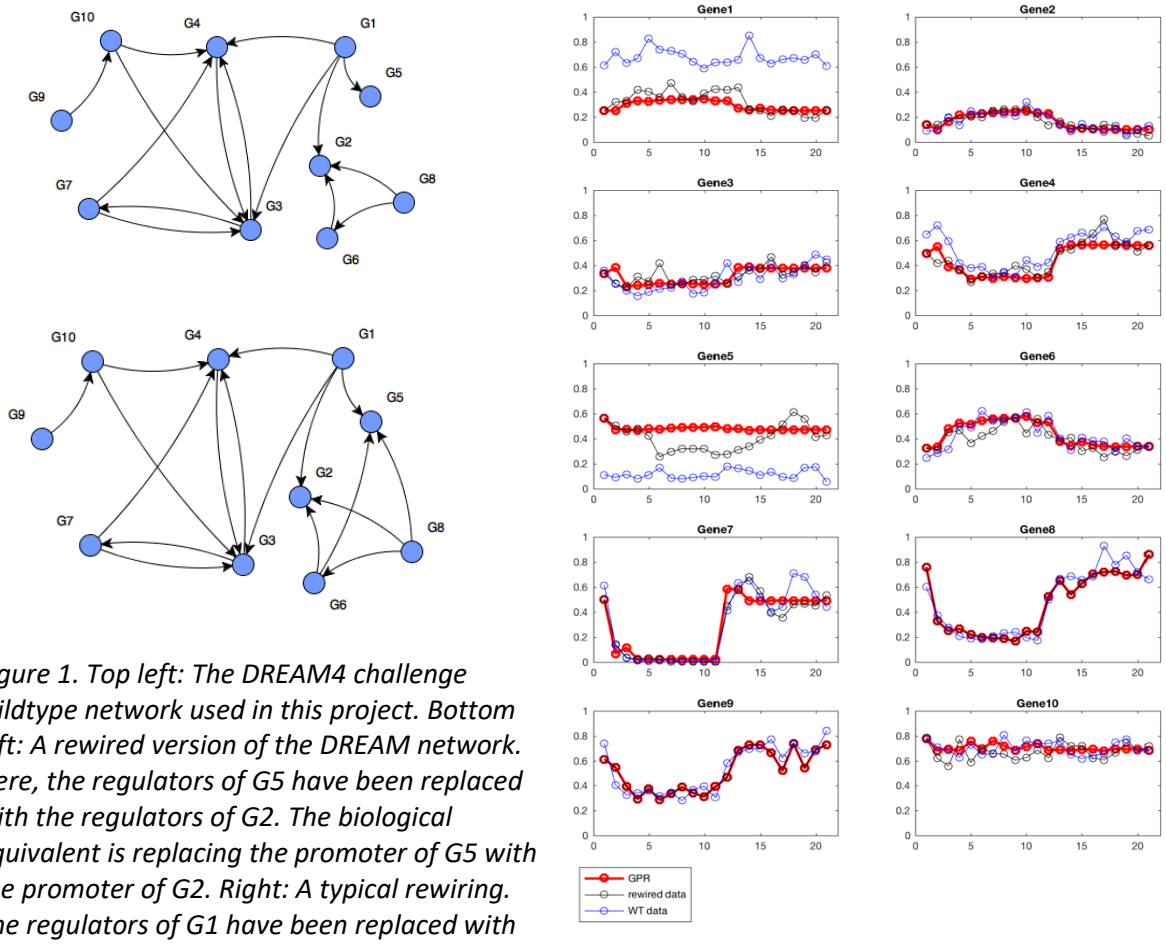

Figure 1. Top left: The DREAM4 challenge wildtype network used in this project. Bottom left: A rewired version of the DREAM network. Here, the regulators of G5 have been replaced with the regulators of G2. The biological equivalent is replacing the promoter of G5 with the promoter of G2. Right: A typical rewiring. The regulators of G1 have been replaced with the regulators of G6. The levels of G8 and G9 are fixed - the GPR is not required to infer these.

We evaluate the predictive accuracy of the model using the mean squared error (MSE). All possible rewirings were simulated, under 5 different perturbations, giving 315 simulations in total. 238 of these simulations (76%) had a MSE smaller than 1. For reference, the MSE in the above figure is 0.98.

Making predictions for network optimisation

In order to obtain a certain phenotype, such as disease resistance, we want to optimise the levels of genes in a network. This is achieved by decreasing the levels of genes that are positive regulators of defence and increasing the levels of genes that negatively impact the defence response. We first test the ability of our models to pick out the rewirings that increase the levels of G5 and decrease the levels of G1 in the DREAM network.

All possible rewirings are simulated and the ones that best fit the optimisation criteria are selected.

In the example in Figure 2, our approach identified the only two single-gene rewirings that resulted in G1 levels decreasing *and* G5 levels increasing *and* the rest of the genes remaining relatively unchanged out of a possible 315. This was achieved by replacing the regulators of G1 with the regulators of either G5 or G6.

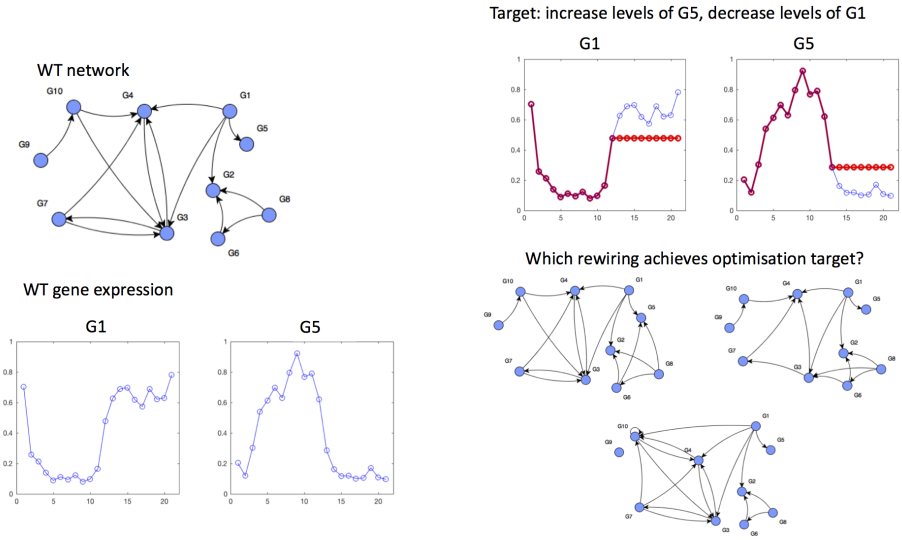

Figure 2: Starting out with just the wildtype network and gene expression data, can GPR predict the impact of rewiring on the network, and pick out the rewirings that optimise the levels of Genes 1 and 5?

Network optimisation via rewiring in Arabidopsis network

A 70-gene network was constructed from Arabidopsis transcription factors using a timeseries containing the gene expression of *Arabidopsis* genes from Windram et al. (2012). These genes were chosen because they had the same response in *Botrytis*-infected leaves and in chitin-treated protoplasts. This is necessary because we plan to use protoplasts as a high-throughput screening method for rewirings that will decreases susceptibility to infection.

Five of the genes in the network have an associated *Botrytis* phenotype.

AT1G02170	G1	AT2G33710	G24	AT3G62420	G47
AT1G09530	G2	AT2G37430	G25	AT3G57480	G48
AT1G19180	G3	AT2G38250	G26	AT3G12910	G49
AT1G22190	G4	AT2G38470	G27	AT3G47500	G50
AT1G25440	G5	AT2G40140	G28	AT3G53600	G51
AT1G25550	G6	AT2G40740	G29	AT4G01250	G52
AT1G27720	G7	AT2G43000	G30	AT4G12040	G53
AT1G27730	G8	AT2G44430	G31	AT4G17500	G54
AT1G28370	G9	AT2G44840	G32	AT4G17880	G55

AT1G42990	G10	AT2G24570	G33	AT4G30080	G56
AT1G54140	G11	AT2G47520	G34	AT4G36990	G57
AT1G62300	G12	AT2G32020	G35	AT4G00270	G58
AT1G67100	G13	AT3G05200	G36	AT4G20380	G59
AT1G70920	G14	AT3G13040	G37	AT5G05610	G60
AT1G71030	G15	AT3G19580	G38	AT5G24110	G61
AT1G71520	G16	AT3G23230	G39	AT5G47230	G62
AT1G72310	G17	AT3G23250	G40	AT5G47370	G63
AT1G61960	G18	AT3G28210	G41	AT5G49520	G64
AT1G80840	G19	AT3G46620	G42	AT5G50570	G65
AT2G03340	G20	AT3G49530	G43	AT5G56960	G66
AT2G04880	G21	AT3G52800	G44	AT5G57660	G67
AT2G23320	G22	AT3G55560	G45	AT5G59820	G68
AT2G24650	G23	AT3G61890	G46	AT5G64340	G69
				AT5G66320	G70

The 70 TFs included in the Arabidopsis network model.

AGI	Gene no.	Gene Name	Phenotype	Source
AT2G03340	G20	*WRKY3*	Positive	Unpublished data
AT2G38470	G27	*WRKY33*	Positive	Unpublished data
AT4G17500	G54	*At-ERF1*	Negative	Foo *et al.,* 2018
AT5G05610	G60	*AL1*	Positive	Unpublished data
AT5G47230	G62	*ERF5*	Positive	Moffat *et al.,* 2012

The genes in the 70 GRN which have an associated phenotype during Botrytis infection. 'Positive' means that the gene helps the Arabidopsis defence response while a 'Negative' phenotype means that the gene hinders the defence response

A GPR model was trained on the expression data for the 70 genes. It then was used to predict gene expression levels in the case of all possible rewirings. An optimisation function was constructed which would have G20, G27, G60 and G62 upregulated, G54 downregulated or unchanged and the rest of the genes in the network unchanged. All the rewirings were then queried to find the ones that fulfilled this criteria best.

Two rewirings fulfilled the criteria for 66 out of 70 genes: replacing the regulatory elements of G16 with the regulatory elements of G23 or G55. In this scenario, the levels of G20 and G60 increase, and the levels of G54 remain unchanged (but do not decrease). In addition, 2 other genes are affected by this rewiring: G16 and G42, whose levels decrease.

Rewiring *in vivo* in Arabidopsis protoplasts

For proof of concept that rewiring can be achieved in Arabidopsis protoplasts, genes that downregulated are rewired with the promoters of highly expressed genes. Protoplasts are then transfected with these constructs and qPCR is used to observe the levels of rewired genes and compare them to the control, which was transformed with a GFP-containing plasmid. The protoplasts are treated with chitin, which is a MAMP present in the cell wall of fungi, and elicits an immune response similar to that of *Botrytis cinerea*.

Two promoter regions upstream of *ZAT11* and *SAP12*, and three gene-coding regions of the genes *AL1, TGA3, WRKY3* and *GFP* were combined to form 8 promoter-gene fusions on a plasmid with Golden Gate.

Arabidopsis protoplasts were then transformed using these rewired constructs. The following day, the protoplasts were induced with chitin for 45 minutes before being frozen in liquid nitrogen.

In order to determine the effect these transformations had on gene expression levels after chitin treatment, qPCR was performed on the protoplast RNA samples. The gene expression of genes *AL1, TGA3, WRKY3* was compared to their expression in the control protoplasts, which were transformed with promoter-GFP fusions. As expected, the rewired genes were upregulated compared to the GFP controls.

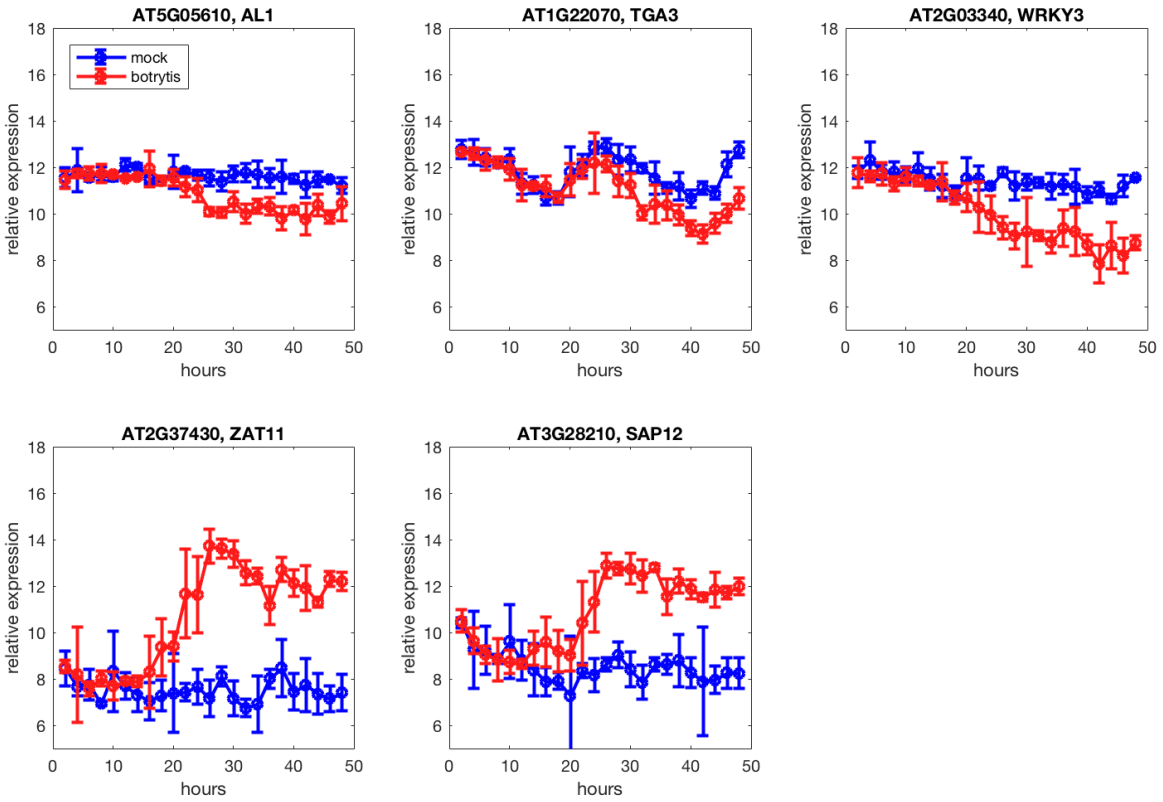

Figure 3. Expression levels of investigated genes in Arabidopsis mock versus infection.

Gene expression over time of the genes used for rewiring. *AL1*, *TGA3* and *WRKY3* are all downregulated in Arabidopsis leaves infected with *Botrytis cinerea* and *ZAT11* and *SAP12* are upregulated. *AL1*, *TGA3* and *WRKY3* are also downregulated in protoplasts treated with chitin. Data from Windram *et al*. (2012).

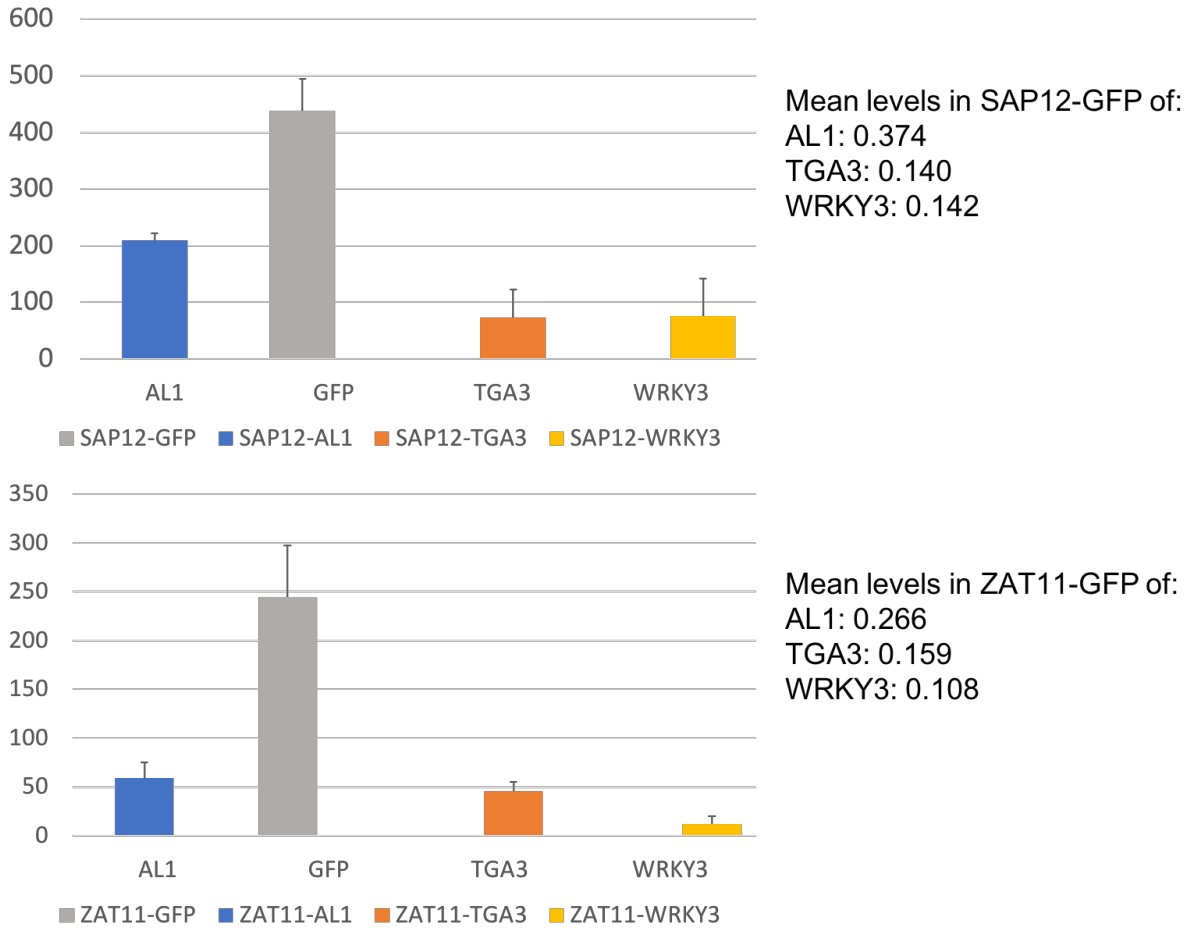

Figure 4. Expression levels as determined through qPCR of rewired genes in protoplasts after treatment with chitin. As expected, all are significantly upregulated compared to their levels in the SAP12/ZAT11-GFP control.

Ongoing work

Now that the proof of concept is demonstrated to be successful rewiring for the GRN70 need to be constructed and trialled in protoplasts and stable transformations in Arabidopsis plants.

2.2. Branching dynamical systems

A key goal in the biological sciences is to identify transcriptional branching. That is, to identify where the expression patterns of key marker genes begin to diverge between two or more conditions. This would allow us to identify the sequence of molecular events that occur during an arbitrary set of perturbations to the system.

We have developed a dynamical systems approach to the inference of transcriptional branching using Gaussian processes. Gaussian processes represent a Bayesian nonparametric approach to regression and classification that is completely defined by a *mean function* and *covariance function*. We have identified how the composition of specific covariance functions can be used to define a branching process of arbitrary complexity i.e., allowing simultaneous comparison of divergence of more than two time-series.

Using this approach allowed us to infer the ordering of branching in *Arabidopsis thaliana* genes following infection with *Pseudomonas syringae DC3000* and a mutant strain *hrpA*, versus mock infection (Penfold, Sybirna et al., Bioinformatics, 2018; manuscript attached).

2.2.1. Branching processes in mammalian systems

During embryonic development, individual cells undertake a series of proliferative or cell fate decisions to eventually form a complete individual, comprised of many cell types. This development epitomises a branching process, and a key aim is therefore to identify the branching structure of individual genes over a range of cell types, which should shed light on the transcriptional programs underpinning individual cell fates. One fate of particular interest is the primordial germ cell, the embryonic precursor of sperm or eggs, that is specified in the early developing embryo following implantation.

Unlike for our plant systems, there is limited time series data available for early mammalian embryogenesis and PGC specification. There does, however, exist collections of single cell RNA-seq data collected over lower resolution time series. By modifying our branching-process dynamical system model, we have developed an approach that simultaneously orders single cell data over a developmental trajectory (pseudotime). By identifying the earliest branching events, we identified that FGF-signalling appeared to be a crucial component of PGC development, which was tested using an in vitro model of germ cell development.

We further identified key enrichment of polycomb repressive complex 1 and 2 and other binding partners of the PRC amongst branching genes. Using our in vitro model, we again verified that inhibition of enzymatic activity of PRC1/2 via small molecules resulted in altered PGC induction rates (Figure 2). Indeed, so thus far, all four of the identified targets have shown a phenotype in PGCLC induction efficiency. This contrasts to around a 1/20 hit rate for random small molecule trials.

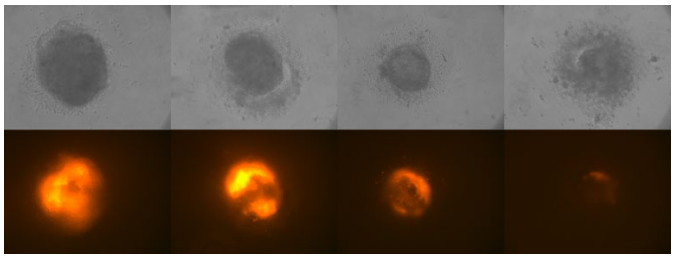

Figure 6. PGCLC-induction from embryoid bodies using a fluorescent reporter for a PGC marker gene. From left to right: control induction, followed by increasing dose of PRC1-enzymatic inhibitor. FACS quantification demonstrates that PGCLC induction efficiency reduces from around 50% to <10%.

These results have formed the basis for a second paper that will be submitted soon (Penfold, Sybirna et al., 2018; manuscript attached).

We demonstrated as a proof of principle the success of our pseudotime approach for time-structured data. Interestingly, we identified targets of a PRC-related gene, MAX, as being highly enriched amongst branched genes. MAX has been demonstrated to play a role in mitotic division, and studies in mouse have shown that knockdown of MAX results in ectopic entry into meiosis. We currently plan to see if rewiring MAX in human PGCLCs can similarly induce early entry into mitotic quiescence or meiosis. For this we will use a proportion of the remaining money to perform preliminary shRNA knockdown studies during PGCLC induction. If successful, we will construct an auxin-inducible degron system for the MAX protein.

Expenditure

Travel	£88.75
Consumables (Related to plant systems)	£1390.20
Consumables (Related to mammal systems)	£1289.13
	£2,768.08

Would you like to claim the £1,000 follow-on fund?
Yes, we would like to claim the £1,000 follow-on fund to be used as described below.

Follow on Plans

Following the proof-of-principle success of our dynamical systems approaches, we have made the code publically available at https://github.com/cap76/BranchingGPs (branching processes), https://github.com/cap76/PGCPseudotime (pseudotime) and https://github.com/cap76/NeRD (dynamical systems). We currently plan to deliver training seminars at the upcoming GARNet meeting. This will require some porting code into a more open environments.

We are currently in the process of testing to see if rewiring identified genes in hESCs can induce meiotic division. For this we will use a proportion of the remaining money to perform preliminary shRNA knockdown studies during PGCLC induction. If successful, we will construct an auxin-inducible degron synthetic system to investigate functional roles.