Nonparametric Bayesian inference of transcriptional branching and recombination identifies regulators of early human germ cell development

Christopher A. Penfold[1,*,‡], Anastasiya Sybirna[1,2,3,*], John Reid[4], Yun Huang[1,3], Lorenz Wernisch[4], Zoubin Ghahramani[5], Murray Grant[6], and M. Azim Surani[1,2,3,§]

[1]Wellcome/CRUK UK Gurdon Institute, University of Cambridge
[2] Wellcome/MRC Cambridge Stem Cell Institute, University of Cambridge
[3]Physiology, Development and Neuroscience Department, University of Cambridge
[4]MRC Biostatistics Unit, University of Cambridge
[5]Department of Engineering, University of Cambridge
[6]School of Life Sciences, University of Warwick
[*]These authors contributed equally
[‡]c.penfold@gurdon.cam.ac.uk, [§]a.surani@gurdon.cam.ac.uk

Abstract

During embryonic development, cells undertake a series of cell fate decisions to form a complete organism, epitomising a branching process; in some instances, these decisions may be reversible, particularly during the onset of disease, exemplifying a recombination process. Single cell transcriptomics provide a rich resource to explore the temporal progression of bifurcations in gene activity, and thus elucidate mechanisms of cell fate decisions. However, identification of transcriptional branching or recombination poses a major statistical challenge. Here, we have developed a comprehensive nonparametric Bayesian approach to the inference of branching and recombination, in the form of branch-recombinant Gaussian Processes (B-RGPs). We use B-RGPs to infer transcriptional branching that occurs during early human development as primordial germ cells (PGCs), the precursors of sperm and egg, are specified in the developing embryo. Using our approach, we identify known master regulators of human PGC development, and predict roles for a variety of signalling pathways and genetic regulators. By concentrating on the earliest signalling events, we identified an antagonistic role for FGF receptor (FGFR) signalling pathway in the acquisition of competence for human PGC fate. Indeed, the experimental validation of our prediction confirmed that pharmacological blocking of FGFR or its downstream effectors (MEK, PI3K and JAK) enhanced the competency for PGC fate *in vitro*. Thus, B-RGPs represent a powerful and flexible data driven approach for dissecting the temporal dynamics of cell fate decisions, providing unique insights into the mechanisms of early embryogenesis.

1 INTRODUCTION

A common class of behaviour encountered in the biological sciences involves branching. In a branching process, often driven by a biological perturbation, a single statistical process undergoes a bifurcation leading to two potentially correlated processes that may, themselves, undergo further branching. Reciprocal behaviour is evident in recombination processes, where two or more statistical processes converge.

A particularly striking example of transcriptional branching occurs in early human embryos, when the inner cell mass (ICM) of the blastocyst segregates into hypoblast and epiblast, with the latter subsequently differentiating into ectoderm, mesoderm, and endoderm in the process of gastrulation (Irie, Tang, and Azim Surani 2014). At around this time, circa weeks 2-3 in humans, the embryonic precursors of gametes, called primordial germ cells (PGC), are also specified (Irie et al. 2015; Kobayashi et al. 2017). Later, at weeks 5-6, specified PGCs undergo comprehensive epigenetic reprogramming that involves the almost complete erasure of DNA methylation marks throughout the genome save for a few escapee regions (Guo et al. 2015; Tang et al. 2015; Gkountela et al. 2015). During this period PGCs also proliferate, and migrate towards the genital ridges where, after colonising the gonads, they begin sexually dimorphic programs of gametogenesis. As germ line precursors, PGCs are ultimately responsible for passing on genetic and epigenetic information to the next generation.

Current models, based on studies in mouse and pig embryos, suggest that two signalling pathways cooperate to allow PGC fate: WNT signalling renders epiblast cells competent to respond to BMP2/4 resulting in the specification of a founder PGC population in the posterior epiblast (Ohinata et al. 2009; Aramaki et al. 2013; Kobayashi et al. 2017). While mouse PGC development has been reasonably well characterised, the specification and development of human PGCs remains only partially understood, primarily due to inaccessibility of early human embryos. Crucially, recent studies have shown that *SOX17* and *PRDM1* are key regulators of PGC specification in humans (Irie et al. 2015; Tang et al. 2015), making their specification distinct from that in mice, which involves the tripartite action of *Prdm1*, *Prdm14*, and *Tfap2c* (Magnusdottir et al. 2013; Nakaki et al. 2013). Due to these differences, along with notable divergence in embryo morphology (Irie, Tang, and Azim Surani 2014), *in vitro* derivation of human PGC-like cells (PGCLCs) from human pluripotent stem cells has emerged as a model to examine the earliest mechanisms regulating hPGC fate (Sasaki et al. 2015; Irie et al. 2015). Interestingly, human ESCs in conventional cultures have low competence to form PGCLCs, but hESCs grown in a specially formulated medium gain competence for PGC fate and respond to BMP signalling giving rise to PGCLCs that presumably resemble pre-migratory *in vivo* PGCs (Irie et al. 2015). However, the mechanisms allowing human PGC competence acquisition remain poorly understood.

System-level measurements of the transcriptome, such as temporally-resolved collections of RNA-sequencing (RNA-seq) data, are increasingly being used to interrogate the mechanisms underpinning the cell fate decisions of PGCs (Guo et al. 2015; Tang et al. 2015; Irie et al. 2015). A key challenge here is the accurate and automated quantification of transcriptional branching dynamics. This remains a difficult task, partially due to datasets being noisy in nature, with missing observations or uneven temporal sampling. The increasing prevalence of single cell transcriptomics provides unique opportunities to further dissect the dynamics of such cell fate decisions. However, these datasets may have low temporal resolution, albeit with observations of many cells, which, if ordered correctly, reflect a continuum of developmental states or responses to perturbations (Yan et al. 2013; Guo et al. 2015; Petropoulos et al. 2016; Borensztein et al. 2017; Huang et al. 2017). Finally, the dynamic nature of different biological systems may also vary significantly, frustrating efforts to find a robust, broadly applicable statistical framework for the inference of branching and recombination.

Nonparametric Bayesian approaches to branching and recombination, such as those based on Gaussian processes regression (Rasmussen and Williams 2006), would therefore be advantageous, providing flexible frameworks able to quantify a broad range of dynamic behaviour with minimal need for parametrisation. These approaches provide robust treatments of uncertainty, and can gracefully cope with missing observations, for example in the case of absence of intermediate cell types. Here we introduce branch-recombinant Gaussian processes (B-RGPs), a comprehensive framework for the inference of arbitrarily complex branching and recombination processes, both in terms of the number of bifurcations, and richness of the underlying dynamics. We first validate our framework in Section 2 using simulated datasets and demonstrate its utility on genome-scale datasets that have previously been used to benchmark two-component branching processes; specifically, we inferred the transcriptional branching and recombination patterns in *Arabidopsis thaliana* microarray time series following inoculation with a bacterial pathogen. In Section 3 we use B-RGPs to dissect the transcriptional program and signalling pathways implicated in human PGC competence and specification using single cell RNA-seq data of PGCs, somatic tissues, pre-implantation embryos, and human embryonic stem cells (hESCs). By adopting an integrative analysis of branching and recombination between PGC and somatic lineages, as well as between *in vivo* and *in vitro* datasets, we correctly identified known PGC genes and putative regulators. In particular, the analysis of earliest branching events highlighted potential roles for FGF receptor (FGFR) signalling in the acquisition of competence for human PGC fate, which was experimentally validated using human *in vitro* models of PGC specification. Indeed, pharmacological blocking of FGFR or its downstream effectors (MEK, PI3K and JAK) enhanced the competency for PGC fate *in vitro*. Thus, our approach can inform genetic and signalling perturbations for cell fate decisions in a time-resolved manner.

2 A COMPREHENSIVE NONPARAMETRIC BAYESIAN APPROACH TO BRANCHING AND RECOMBINANTION

Within a Bayesian setting, Gaussian processes (GPs) can be used to represent prior distributions over functions, providing a flexible framework for regression that is naturally adept to deal with uncertainty and missing data (Rasmussen and Williams 2006). This makes GP-based approaches ideal frameworks for quantifying the dynamics of gene expression from biological observations (Stegle et al. 2010; Kalaitzis and Lawrence 2011; Breeze et al. 2011; Hensman, Lawrence, and Rattray 2013), with GP-based approaches increasingly being used to interrogate complex single-cell datasets (Reid and Wernisch 2016; Campbell and Yau 2015; Buettner et al. 2015; Zwiessele and Lawrence 2016). Previous studies by Yang et al. (2016) have demonstrated that GPs can be used to explicitly quantify the dynamics of branching processes, but are only applicable for data with a single bifurcation. Here we have developed branch-recombinant Gaussian Processes (B-RGPs, Supplementary Materials Section 1), which allow for the inference of arbitrarily complex branching and recombination processes, both in terms of the number of bifurcations, and the underlying complexity in the dynamics of individual branches. B-RGPs were encoded in MATLAB using the gpml package (Rasmussen and Williams 2006), and code is available to download from GitHub: https://github.com/cap76/BranchingGPs.

As a preliminary test of the B-RGP framework we fitted a variety of simulated time series datasets, with the results summarised in Supplementary Materials Section 2.1. In general, our methodology shows increased predictive accuracy over independent GPs when the underlying

data were branch-recombinant, but was also robust to model mismatch i.e., accuracy was comparable to independent GPs when the datasets were non-branching in nature.

To demonstrate the utility of B-RGPs on genome-scale applications, we inferred the transcriptional branching structures that exist in the model plant *Arabidopsis thaliana* following inoculation with a bacterial pathogen *Pseudomonas syringae* pv. tomato DC3000, and a disarmed mutant strain hrpA (Lewis et al. (2015); GEO GSE56094). Whilst the DC3000 variant delivers 28 effector proteins into the plant cell to subvert the plants immune response, the hrpA mutant lacks the apparatus necessary for doing so, thus reporting a classical immune response when compared to mock infected leaves. Comparison of the hrpA-inoculated time series with the DC3000-inoculated time series has previously been used a test case for two-component branching processes {Yang, 2016 #70}, providing useful insights into the development of disease. This dataset therefore provides an obvious biological benchmark for our method. Here we enumerated over a number of possible 0-branch, 1-branch and 2-branch processes using B-RGPs, optimising hyperparameters to their MAP values on a gene-by-gene basis, and using the Bayesian Information Criterion (BIC) to assign a branch structure to each gene (see Supplementary Materials Section 2). This analysis grouped genes into six groups, with example expression patterns shown in Figure 1(a). These groups are summarised as follows: hrpA branches from the control, with DC3000 branching from hrpA at a later point (Group 1), or hrpA and DC3000 independently branch from the control (Group 2), which collectively represent the immune response genes targeted by effectors; DC3000, but not hrpA, branches from control (Group 3), representing host susceptibility genes that have been targeted by effectors; hrpA, but not DC3000, branches from the control (Group 4), likely representing immune genes that have been targeted by effectors prior to their natural immune response times; both DC3000 and hrpA jointly branch from the control, but not from one another (Group 5), representing core immune response genes not targeted by effectors; no branching exists (Group 6), representing genes unaffected by plant immunity or pathogen virulence strategies. These individual groups were appropriately enriched for GO terms (Supplementary Table 1) and for known targets of pathogen effectors identified in Y1H studies (Mukhtar et al. (2011); Figure 1(c, d, e); see Supplementary Section 2.2).

Finally, to test the usefulness of having an estimated time of branching, we checked if branch times were predictive of direct and indirect targets of key regulators. Here we focused on WRKY11 and WRKY17. WRKY11/17 are amongst the earliest branching transcription factors (TFs) and have previously been implicated in the *Arabidopsis* response to *P. syringae* (Journot-Catalino et al. 2006). Both genes showed branching between control and hrpA, and between hrpA and DC3000 consistent with (i) their immune responsive expression and (ii) their manipulation by effectors. Genes that branched between control and hrpA and between hrpA and DC3000 were assigned a Euclidean distance based on the position of their branch times from that of WRKY11 or WRKY17. We then compared the distributions of these Euclidean distances for the subset of genes identified as being differentially expressed (DE) in knockout mutants of WRKY11/17 (Journot-Catalino et al. 2006) versus the distribution of the subset of genes that were not differentially expressed (NDE) in those mutants. Our results show that DE genes had significantly smaller Euclidean distances than NDE genes ($p<0.05$ for WRKY11 and $p<0.005$ for WRKY17 using two-sided Student's t-test; Figure 1(f, g)), suggesting that genes that branched at similar times to WRKY11/17 were more likely to also be direct or indirect targets. WRKY11/17 are TFs and could exert direct regulation of their targets by binding to their regulatory elements.

To check this, we searched for the presence of WRKY motifs within a 1kb promoter region using FIMO (Grant, Bailey, and Noble 2011); specifically, the stringent WRKY binding site (Wbox) motif, TWGTTGACYWWWW, identified by Ciolkowski et al. (2008). Here, we looked at the frequency of Wbox motifs ($p<0.0001$) in sets of genes whose branch times were increasingly distal from WRKY11. These groups were based on: (i) genes whose Euclidean distance $d<1$, representing the closest 156 genes; (ii) genes whose Euclidean distance $d<2$, representing the closest 454 genes; and (iii) the closest 2000 genes. As positive and negative controls, we also included the 157 genes that were DE in the WRKY11 KO, and 2000 genes randomly selected from Group 6 (genes with no branching). Our results showed a clear trend of increasing frequency of Wbox motifs in sets of genes whose branch times were closest to that of WRKY11 (Figure 1(h); see also Supplementary Table 3). Together, these results suggest that estimation of branch times may be useful for identifying direct and indirect targets of perturbed genes.

Together, these results collectively demonstrate the efficacy of B-RGPs, and highlight their use in extracting temporally resolved information from complex biological datasets.

3 TRANSCRIPTIONAL BRANCHING AND RECOMBINATION IN HUMAN PRIMORDIAL GERM CELLS

Early human PGC development remains underexplored, since the use of human embryos at the time of PGC specification is impracticable. On the other hand, differences between human and rodent development have necessitated the use of human-specific model systems (Irie, Tang, and Azim Surani 2014). As a consequence, statistical approaches are required to correctly leverage multiple *in vivo* and *in vitro* datasets to identify priority targets for more focused experiments.

Here we have used B-RGPs to infer the branching dynamics of early human PGCs to identify the earliest regulators of PGC fate. For this we used single cell RNA-seq data from hPGCs and age-matched somatic tissues from weeks 4 through to 19 (Guo et al. (2015); GEO GSE63818), as well as pre-implantation embryos (from oocytes to blastocysts), and conventional hESCs at passages 0 and 10 (Yan et al. (2013); GEO GSE36552). Together, these datasets provide unique opportunities for investigating the dynamics of decision making in early human embryos, as cells undergo specification towards the germ line, and provide a rich testbed for addressing the differences between the *in vivo* dynamics of PGC development, and the dynamics reflected in pseudo-temporally ordered ESCs. Whilst there is currently no single-cell transcriptomic data available for competent hESCs and PGCLCs, the ability of GPs to cope with missing and unevenly sampled data should allow useful insights to be extracted from conventional hESC and *in vivo* hPGC data.

To separate out the developmental trajectories of PGCs from those of gonadal somatic cells we initially ordered single cells along a two-component Matérn branching process using a Gaussian process latent variable model (GPLVM; Lawrence (2003, 2005)) with informative priors placed over the pseudotimes centred on the cells' developmental stages (Reid and Wernisch 2016). Preliminary benchmarking of branch-recombinant GPLVMs suggested that cells could be accurately (pseudo-temporally) ordered along a branching process using around 40 genes (see Supplementary Materials Section 3), and we therefore used the expression of 44 marker genes identified in Irie et al. (2015), which included PGC, pluripotency, mesoderm, and endoderm

markers. We evaluated the accuracy of our approach using unlabelled data, assuming that the branch-label and capture-times for hESC and blastocyst cells were unknown variables to be inferred within the algorithm (see Supplementary Section 4). Our inference correctly placed the unlabelled late-blastocyst cells between morula-stage and week four cells (Figure 2(a)). Furthermore, whilst hESCs are derived from the ICM of blastocysts, our approach placed them between day 6 blastocysts and 4-week-old cells (PGCs and soma), albeit with some degree of overlap. This suggests that hESCs cultured under these conditions are developmentally more advanced than cells of the ICM, consistent with recent suggestions that place hESCs roughly in line with post-implantation embryos (Tang et al. (2016); Figure 2(a), inset). Finally, it is also worth noting that our inference assigned later pseudotimes to hESCs at passage 10 compared to passage 0 ($p<0.05$, two-sided t-test; Figure 2(b)), potentially reflecting culture-induced adaptations.

Once the initial ordering of cells had been established, we searched for more complex branching on a genome-wide scale, in order to quantify the potential divergence between ESCs and inferred *in vivo* dynamics. To do so, we fixed pseudotimes based on our preliminary analyses, and optimised hyperparameters on a gene-by-gene basis (see Supplementary Materials Section 4). The developmental trajectory of PGCs was assumed to represented the basal process, due to this class having the most data points, and we simultaneously searched for bifurcations between hPGCs and somatic cells, and between inferred *in vivo* dynamics of PGC specification and the likely *in vitro* dynamics required to drive ESCs towards week 4 PGC fate. For the latter case, we were specifically interested in interpolating the trajectory required to drive the transcriptome of ESCs towards week 4 PGCs, and therefore assumed an explicit branch-recombinant process models (for full details see Supplementary Section 6). In total, we therefore considered three branch-recombinant structures: (i) somatic cells branched from the base process (hPGCs), with hESCs and PGCs following an identical process (see Figure 2(c, d, e)); (ii) both soma cells and hESCs branched from the main process, with hESCs later recombining (Figure 2(e)); and (iii) all processes were identically distributed.

3.1 In vivo dynamics of human PGC development.

Using B-RGPs we identified genes that branched between hPGCs and human somatic cells. In total, 3,930 genes were identified as being up-regulated in hPGCs versus soma, with 1,355 down-regulated. It is likely that these genes are involved in a broad range of biological processes, including the acquisition of competence for PGC fate, PGC specification and maintenance, as well as epigenetic reprogramming, migration, and gametogenesis. We therefore performed a preliminary GO and KEGG pathway analysis using a permissive p-value ($p<0.1$, hypergeometric test), with the aim of identifying biological processes that branched early in development. We subsequently ordered terms according to when 50% of their associated genes had branched, as in Yang et al. (2016), with a heatmap representation of this temporal progression shown in Supplementary Figure 11 (see also associated Supplementary Table 4). Our analysis revealed several early terms related to BMP and WNT signalling, with subsequent terms related to epigenetic reprogramming, proliferation and cell migration, and later terms relating to testosterone signalling, meiosis, and gametogenesis, reflecting the expected progression of hPGC development (Lawson et al. 1999; Ohinata et al. 2009; De Felici 2013; Leitch, Tang, and Surani 2013; Kobayashi et al. 2017).

A histogram of the time of branching between PGCs and soma (Figure 2(c)) shows a multi-modal distribution, with most responses occurring after blastocyst-stage, but prior to week 4, consistent with the expected timing for PGC specification at around weeks 2-3 of development (see e.g., Tang et al. (2016)). Importantly, two key regulators of human PGC fate, *SOX17* and *PRDM1* (Irie et al. 2015), were shown to branch early in the pseudotime series, prior to late PGC markers such as *DAZL* (see Figure 2(c, d, e)). In addition to regulating human, but not mouse, PGC fate, *SOX17* is a known endoderm TF in mice and humans (Kanai-Azuma et al. 2002; Seguin et al. 2008). It is therefore possible that a subset of its target genes are shared between human endoderm and PGC lineages. To test this, we looked at genes that were differentially expressed in hESCs following overexpression of *SOX17* compared to parental hESC lines (Seguin et al. (2008); GSE10809). Genes that branched between PGCs and soma were statistically enriched for genes that were differentially expressed in response to *SOX17* overexpression ($p<1x10^{-35}$, hypergeometric test using Enrichr). Reanalysis of this microarray data shows that common genes included the known PGC regulator *PRDM1*. Crucially, the overlapping targets of *SOX17* were, themselves, shown to branch earlier than branching times in the dataset as a whole ($p<1x10^{-7}$, Kolmogorov–Smirnov test), and earlier then testosterone signalling, known to be involved in later male germ cell development (Figure 2(g)). Together, these results highlight the crucial early role for *SOX17* in hPGC fate decision.

To identify additional regulatory mechanisms that could be contributing to PGC specification and maintenance, we used Enrichr (Chen et al. 2013; Kuleshov et al. 2016) to find transcriptional signatures present within our set of branched genes, and a Kolmogorov-Smirnov (KS) test to identify those that were significantly earlier than branching times as a whole. That is, we looked for early and statistically significant overlaps between genes that branched (PGCs versus soma) and those that were up-regulated or down-regulated in the literature in a variety of species and cell types following knockout or overexpression studies, or upon treatment with hormones, growth factors or cytokines. Although none of the perturbations were performed directly in PGCs, and results might therefore be contextually very different to our system, the presence of an early and statistically robust transcriptional signature is worthy of further investigation. Furthermore, early enriched signatures, should highlight key signalling molecules that are secreted from a supporting niche in a non-autocrine fashion.

In agreement with the early branching of genes with BMP-related GO/KEGG terms around the time of PGC specification, comparison with the Enrichr database showed a strong overlap of branched genes with those differentially expressed in hESCs following treatment with BMP4 ($p<5x10^{-40}$, Enrichr). The identified targets of BMP4 branched early in the pseudotime series compared to branch times overall ($p<5x10^{-4}$, KS-test; Figure 2(h)), reflecting the known role of BMP4 as a PGC-inductive signal (Lawson et al. 1999). We further identified the transcriptional signatures of several putative BMP and WNT effector genes, including Parathyroid Hormone Like Hormone (*PTHLH*, $p<5e^{-15}$, Enrichr; $p<0.1$, KS-test), which increases mesenchymal cells' responsiveness to BMP4 (Hens et al. 2009), and has been implicated in the emergence of germ cell cancers (Sandberg, Meloni, and Suijkerbuijk 1996; Mostert et al. 1998). Other signatures included X-linked inhibitor of apoptosis (*XIAP*, $p<5x10^{-31}$, Enrichr; $p<5x10^{-4}$, KS-test), which, besides modulating BMP and WNT signalling, has also been associated with seminomas, cancers originating from PGCs (Kempkensteffen et al. 2007). Within branched genes, we also noted strong signatures of genes involved in canonical WNT signalling, including both *CTNNB1*

($p<5x10^{-18}$, Enrichr; $p<0.01$, KS-test), *GSK3β* ($p<5x10^{-3}$, Enrichr; $p<0.1$, KS-test), as well APC Membrane Recruitment Protein 1 (*AMER1,* p<5e-16, Enrichr; $p<5x10^{-4}$, KS-test), and its target, Wilms Tumor 1 (*WT1*, $p<5x10^{-13}$, Enrichr; $p<0.05$, KS-test). These findings are in keeping with WNT and BMP serving as major PGC induction signals and thus prove the efficacy of our framework.

Amongst the identified enriched gene signatures was overexpression of the transcription factor *NANOG* (Enrichr, $p<1x10^{-6}$; $p<0.05$ KS-test). In mouse *in vitro* models NANOG is sufficient to directly induce PGC-like cell fate by binding and activating enhancers of *Prdm1* and *Prdm14* (Murakami et al. 2016), although loss of function studies have yielded equivocal results as to its requirement for mouse PGC fate (Chambers et al. 2007; Yamaguchi et al. 2009; Carter et al. 2014). NANOG is highly expressed in human PGCs and PGCLCs, but, unlike in mouse cells, its overexpression does not induce PGCLCs (Kobayashi et al. 2017). Nevertheless, we identified a number of putative NANOG binding sites within the 10kb flanking regions of *PRDM1* using FIMO (Supplementary Table 6), suggesting a possible functional roles. It would be of interest to test the involvement of NANOG in hPGCLC fate using an inducible knockout.

Branched genes were also highly enriched for the signature of epigenetic and chromatin modifiers, including the histone demethylase *KDM3A* ($p<1x10^{-11}$, Enrichr; $p<0.01$, KS-test), methyltransferase *DNMT1* ($p<1x10^{-6}$, Enrichr; NS, KS-test), and dioxygenases *TET1 (*$p<1x10^{-4}$, Enrichr; NS, KS-test)*, TET2* ($p<1x10^{-4}$, Enrichr; NS, KS-test), and *TET3* ($p<1x10^{-6}$; NS, KS-test), which contribute to PGC DNA demethylation via conversion of the modified genomic base 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC) (Ito et al. 2010). In general, genes associated with perturbations of epigenetic and chromatin modifiers appeared to branch later than genes associated with perturbations of WNT and BMP signalling.

In addition to picking up known PGC regulators, we were also able to identify pathways potentially involved in human PGC fate, including a putative role for tumour necrosis factor (TNF) signalling in PGC development, based on enrichment of the KEGG term 'TNF signalling pathway', and the TNF transcriptional signature ($p<5x10^{-13}$, Enrichr; $p<0.05$, KS-test). Targets of TNF were shown to branch later than BMP4/SOX17-responsive genes, and around the same time as genes that branched in response perturbations of *TET1*, *TET2*, and *TET3* (Figure 2(i)). This places the timing of TNF-signalling roughly in concordance with epigenetic reprogramming in proliferating, migratory PGCs. Although TNFA has roles in apoptosis, previous studies in mouse models suggest it can stimulate proliferation of pre-gonadal PGCs in culture (Kawase et al. 1994; Makoolati, Movahedin, and Forouzandeh-Moghadam 2016). Whilst *TNF,* itself, was not identified amongst the branched genes, and its expression appears to be generally low in PGCs (Tang et al. 2015), one of its receptors, *TNFRSF1A*, was found to be branched, and it may be that TNF is secreted from surrounding tissues. Indeed, *TNFA* is expressed in Schwann cells (Wagner and Myers 1996), and TNF signalling might therefore support PGCs proliferation and survival as they migrate along nerve fibres towards the genital ridges (Mollgard et al. 2010).

3.2 Differences between inferred in vivo and in vitro dynamics of human development

Current knowledge of human germ cell specification is largely informed by *in vitro* PGCLC derivation from hESCs. While it is challenging to directly validate these findings *in vivo*, we

reasoned that we could highlight the most relevant processes by comparing the inferred dynamics of PGC induction in early embryos versus pseudo-temporally ordered hESCs.

To this end, we looked at genes whose expression branched between hESCs and the inferred *in vivo* dynamics of PGC development. By assuming a branch-recombinant process we were able to identify the initial bifurcation, but also interpolated the likely dynamics required to drive ESCs towards PGC fate. This analysis identified 1,331 genes, the majority of which were up-regulated in hESCs compared to *in vivo*. The timing of divergence between inferred *in vivo* and *in vitro* dynamics is shown in Supplementary Figure 10(c). GO analysis of these groups identified a number of enriched terms, including those relating to cell-cell adhesion, response to hormones, and, importantly, response to BMP4 and WNT, as well as terms relating to germ cell development and meiosis (see Supplementary Table 5).

Amongst the most highly enriched transcriptional signatures was the overexpression of *SOX17* ($p<1x10^{-13}$, Enrichr), a critical regulator of PGC fate. *PRDM1*, another key regulator of PGC fate, was not identified amongst the enriched lists due to a lack of human *PRDM1* perturbations in the Enrichr database, however *de novo* motif analysis using DREME (Bailey 2011) on 1kb windows upstream of the TSS identified two motifs ($p<1x10^{-17}$, $p<5x10^{-7}$) resembling that of human *PRDM1* ($p<0.01$, $p<5x10^{-3}$, TOMTOM, Gupta et al. (2007)). Manual comparison against bulk RNA-seq data of *PRDM1*-overexpressing cells (Kobayashi et al. 2017) showed a significant overlap ($p<5x10^{-24}$, hypergeometric test). Together these results highlight the importance of *SOX17/PRDM1* in the specification of hPGCLCs from hESCs, with recent studies demonstrating that the concomitant induction of *SOX17* and *PRDM1* expression in hESCs is sufficient to drive PGCLC fate, even in the absence of cytokines (Kobayashi et al. 2017).

A strong signature was also noted for *KIT* ($p<1x10^{-6}$, Enrichr), highlighting a potentially important role of its ligand, SCF, in PGC cell fate. SCF has been implicated in PGC proliferation and migration (Hoyer, Byskov, and Mollgard 2005; Mollgard et al. 2010) and represents a component of *in vitro* culture conditions used to derive PGCLCs from ESCs (Irie et al. 2015). Although *KITLG* is not expressed in hPGCs or hPGCLCs (Tang et al. 2015; Irie et al. 2015), the expression of *KIT* is significantly up-regulated in hPGCs, and SCF (encoded by *KITLG*) is potentially secreted from adjacent cells, with SCF expressed in Schwann cells (Mollgard et al. 2010), as well as Sertoli and Leydig cells (Sandlow et al. 1996).

De novo motif analysis using DREME further identified several SOX motifs, and the most highly enriched transcriptional signature from Enrichr was for a knockout of *SOX2* ($p<5x10^{-23}$). *SOX2* is a core pluripotency factor in mouse and human, and is also involved in neural ectodermal differentiation (Masui et al. 2007; Fong, Hohenstein, and Donovan 2008; Thomson et al. 2011). Interestingly, while it is required for mouse PGC fate (Campolo et al. 2013), it is suppressed in human PGCs and PGCLCs (Tang et al. 2015; Irie et al. 2015). Thus, using our approach, we are able to pick up human-specific PGC transcriptional signatures.

The enrichment of a *SOX2* KO signature may indirectly reflect the up-regulation of *PRDM1*, a known repressor of *SOX2* (Lin et al. 2014). However, recent studies have highlighted the importance of the switch between SOX2/OCT4 to SOX17/OCT4 in mouse endodermal specification (Aksoy et al. 2013), and the strong signature identified here points to the importance of correctly regulating *SOX2*/*SOX17* in PGC development. It is therefore possible

that timely *SOX2* downregulation allows *SOX17*-driven PGC specification. Indeed, overexpression of *SOX2* in hESCs under conditions favouring germline differentiation skewed cell fate towards neural lineages (Lin et al. 2014), suggesting a potential competition between SOX2 and SOX17 in this context.

3.3. Early branching events identify key regulators of competence for PGC fate

hESCs in conventional cultures (KSR-based medium supplemented with FGF2) have low potential to form germ cells, but acquire competency for PGCLC fate in the presence of TGFβ, LIF, FGF2 and the inhibitors of GSK3β (CH), p38 (SB), JNK (SP) and MEK (PD03) kinases (4i medium; Gafni et al. (2013); Irie et al. (2015)). 4i hESCs self-renew in a competent state and can form high numbers of PGCLCs when exposed to BMP2/4 and supporting cytokines in embryoid body (EB) cultures (Irie et al. 2015). The mechanisms underlying such dramatic change in developmental potential of 4i (competent hESCs) versus conventional hESCs remain to be elucidated.

Since B-RGPs provide a high temporal resolution, we decided to focus on signalling that may be involved in conferring PGC competence to hESCs. For this, we first compared the genes up-regulated in hESCs to the inferred *in vivo* dynamics for PGCs. This revealed high enrichment for the downregulation of BRAF ($p<5x10^{-7}$, Enrichr; $p<1x10^{-4}$, KS-test) and FGF2 ($p<1x10^{-15}$, Enrichr), upstream regulators of MEK signalling. Crucially, FGF2-responsive genes were found to branch very early *in vivo* (soma vs PGCs) compared to overall branch times ($p<5x10^{-5}$, KS-test). This branching occurred prior to the BMP4/SOX17-responsive genes ($p<0.05$, FGF2 vs BMP4; NS, FGF2 vs SOX17, see Figure 2(h)), suggesting that, FGF2 may function earlier than key specifiers of PGC fate, with potential roles in conferring competency for PGCLC fate. While FGF2 is present in both conventional and 4i conditions, 4i cells are maintained in the presence of an inhibitor of MEK, an effector of FGF receptor (FGFR) signalling. It is therefore possible that partially negating the effect of FGF signalling via the inhibition of downstream pathways of FGFR, might contribute to PGC competence acquisition by pluripotent cells.

Comparison of branching between soma and PGCs identified early signatures for a number of kinase perturbations that mimic the use of the inhibitors in the 4i medium sustaining competent hESCs. These included: (i) knockdown of GSK3B ($p<1x10^{-7}$, Enrichr; $p<0.05$, KS-test) (CH is a GSK3β inhibitor); (ii) knockdown of MAP2K1 ($p<5x10^{-8}$, Enrichr; $p<0.005$, KS-test), a component of the MAP kinase signal transduction pathway upstream of MEK signalling (PD03 is a MEK inhibitor); and (iii) MAPK14 ($1x10^{-8}$, Enrichr; $p<0.01$, KS-test), a regulator of p38 (SB is a p38 inhibitor). Branch times for genes associated with these perturbations were thus suggestive of early roles. Of note, we did not identify statistically significant overlaps with pathways downstream of JNK kinases, inhibited by the fourth inhibitor in 4i hESCs (SP).

In order to experimentally test the involvement of these signalling pathways in PGC competence, we compared PGCLC induction efficiencies from hESCs cultured in either complete competent (4i) medium (control) or lacking one of the 4 inhibitors: (i) "4i-CH" (4i without GSK3B inhibitor); (ii) "4i-PD03" (4i without MEK inhibitor); (iii) "4i-SB" (4i without p38 inhibitor), and (iv) "4i-SP" (4i without JNK inhibitor). hESCs were cultured in these alternate media for at least one passage and collected for standard PGCLC differentiation (as in Irie et al. (2015)). PGCLC induction efficiency was then quantified by flow cytometry (Figure 3). This was

facilitated by the use of PGC-specific knockin reporter cell line NANOS3-tdTomato (N3tdTom; Kobayashi et al. (2017)) and staining for PGC surface markers AP and CD38 (Irie et al. 2015).

This identified that while removal of CH (GSK3βi) and SP (JNKi) from 4i did not significantly change PGCLC competence, withdrawal of PD03 (MEKi) and SB (p38i) strongly reduced the propensity of hESCs to form PGCLCs (Figure 4(a)), consistent with a predicted role for MEK and p38 but not JNK signalling in conferring PGC competence. Considering the crucial role of WNT signalling for PGC competence in mice (Ohinata et al. 2009; Aramaki et al. 2013), it was surprising that CH (WNT agonist, Ying et al. (2008)) withdrawal did not impact on PGCLC competence. We hypothesized that 4i hESCs might produce WNT in an autocrine fashion. Indeed, the number of PGCLCs generated from hESCs grown with WNT pathway inhibitor IWR-1 (IWR, Chen et al. (2009)) either in the presence or the absence of CH, was reduced compared to the control (Figure 4(b)).

Notably, "4i-CH" and "4i-SB" cultures deteriorated at later (>4) passages, reflecting the importance of these inhibitors to sustain self-renewal of 4i hESCs (Gafni et al. 2013). In particular, low competency observed upon withdrawal of SB (p38i) was likely caused by the induction of trophectoderm-like differentiation, as judged by the expression of lineage markers, *CDX2* and *HAND1* (Supplementary Figure 13).

The most drastic effect on PGCLC competence was observed upon withdrawal of MEK inhibitor, PD03, with a strong reduction in the induction of PGCLC (Figure 5(b)), although unlike "4i-CH" or "4i-SB", "4i-PD03" cells could be maintained for many passages (>20). These cells, however, displayed a much flatter morphology than 4i hESCs and expressed slightly lower levels of *POU5F1* (*OCT4*), and variable levels of *SOX2* (Supplementary Figure 14). Importantly, the defect of "4i-PD03" hESCs to differentiate to PGCLCs could be rescued (within one passage) if PD03 was reintroduced (Figure 5(c)).

Since our analysis revealed early branching of an upstream component of this signaling pathway, namely FGF2, we asked if inhibition of FGF receptor (FGFR) could mimic the effect of MEK inhibition. To address this, we transferred the cells cultured in "4i-PD03" to "4i-PD03+PD17" (4i where MEK inhibitor PD0325901 was substituted for FGFR inhibitor PD173074). Not only did FGFR inhibition rescue the differentiation defect of "4i-PD03" cells, but it also markedly enhanced PGCLC competence compared to the 4i control (in the presence of high PD17 concentrations; Figure 5(d)). However, these cells showed decreased viability and collapsed after 2-3 passages. Importantly, FGFRi showed a dose-dependent positive effect on competence (Figure 5(d)) and cells grown with lower PD17 concentration could be maintained longer. Together, these data show that, in agreement with our computational prediction, FGF-MEK pathway is a negative regulator of human PGCLC competence; blocking this cascade by FGFRi or MEKi promotes PGCLC competence in hESCs.

Next, we explored the reason for enhanced competence of FGFRi-treated versus MEKi-treated cells. Major downstream components of FGFR signalling are RAS-MEK-ERK, PI3K-AKT, and JAK-STAT (Lanner and Rossant 2010). Indeed, the comparison of dynamics of PGC development with somatic cells identified strong enrichment for these three pathways. Thus, we saw enrichment of appropriate GO/KEGG terms, and transcriptional signatures of EGFR ($p<5x10^{-29}$) and Platelet Derived Growth Factor Receptors Alpha (PDGFRA, $p<1x10^{-13}$, Enrichr)

and Beta (PDGFRB, $p<5x10^{-7}$, Enrichr), implicated in the activation of ERK, AKT, and STAT1/3/5 signaling pathways, as well as *MET* ($p<5x10^{-7}$, Enrichr), which functions upstream of RAS-ERK and PI3K-AKT signalling. We also noted a strong signature for the down-regulation of JAK1/2 ($p<5x10^{-20}$), and an enrichment for PI3K signalling, which included transcriptional signatures of *AKT1* ($p<5x10^{-11}$, Enrichr), *CSF1R* ($p<5x10^{-5}$, Enrichr), the regulatory subunit of PI3K leading to activation of the AKT1, and *NTRK3* ($p<5x10^{-5}$), which also activates PI3K-AKT signalling. PI3K/AKT pathway was of particular interest, since it is activated downstream of insulin signalling, which was one of the most statistically enriched ligand perturbations from the Enrichr database ($p<1x10^{-9}$, Enrichr) promoting ESC to PGC transcriptome transition.

We therefore hypothesized that simultaneous inhibition of MEK and other pathways downstream of FGFR could mimic the effect of FGFRi and increase PGCLC competence. To test this, we supplemented complete 4i cultures with either LY (LY294002, PI3K inhibitor) or CAS (CAS457081-03-7, JAK inhibitor). This revealed a trend for enhanced PGCLC competence of "4i+LY" hESCs (Figure 6(a)). Importantly, "4i+LY" cells did not exhibit increased cell death as seen with FGFRi; instead they proliferated at rates similar to 4i hESCs and formed homogeneous dome-shaped colonies. Furthermore, "4i-PD03+LY" cells, especially at later passages exhibited competence comparable to control 4i hESCs (Figure 6(a)), pointing to a synergistic action of MEK/ERK and PI3K/AKT signalling pathways. We also observed a positive effect of JAK pathway inhibition on PGCLC competence (Figure 7(b)). This is in line with the enhanced competence of hESCs cultured in 4i without LIF, an agonist of JAK-STAT (Figure 6(c)). Interestingly, differentiation of 4i+CAS" hESCs yielded more $CD38^{+}$ cells (Figure 7(c)), which represent more mature PGCLCs (Irie et al. 2015). Of note, longer culture of hESCs with JAK inhibitor changed colony morphology and subsequently reduced competence (Figure 6(b)). These cultures could not be maintained, highlighting differential requirements for JAK/STAT signalling in 4i versus conventional hESCs (Gafni et al. 2013; Onishi and Zandstra 2015). It is therefore possible to speculate that the observed effect of FGFR inhibition could be explained by: (i) enhanced competence due to PI3K pathway inhibition and transient JAK/STAT inhibition; (ii) decreased viability and loss of pluripotency due to JAK/STAT inhibition. The relationship between these signalling pathways and their contribution to pluripotency maintenance and PGC competence acquisition warrant further investigation. Together, these data show that B-RGPs could predict the relevance of specific signalling pathways to PGC competence even in the absence of all relevant data points (single-cell transcriptomes of 4i (competent) hESCs and PGCLCs).

Finally, we set out to test if in addition to promoting competence, blocking MEK signalling could also enhance PGCLC specification. To this end, we supplemented the differentiation medium with MEK inhibitor PD03 or induced a genetic construct for dominant-negative ERK expression (DN-ERK) at the onset of differentiation (Figure 7(a)). Intriguingly, both perturbations reduced PGCLC numbers compared to controls (Figure 7(b,c)). ERK activation during PGCLC induction is likely triggered by the supplementation of the differentiation medium with EGF (Irie *et al*., 2015), which activates similar intracellular pathways to FGF ligands (Schlessinger 2004). Of note, inhibition of PI3K throughout differentiation did not abrogate PGCLC specification, suggesting that, unlike MEK, PI3K signalling is dispensable for PGCLC induction (Supplementary Figure 15). Altogether, this shows that the timing of FGFR

signalling is crucial for context-specific cell fate decisions and our methodology allows accurate timing predictions in developmental systems.

4 DISCUSSION

Transcriptional branching and recombination are frequently encountered in developmental biology and upon system perturbations. Differentiation of stem and progenitor cells to diverse terminal cell fates can be regarded as an example of branching. On the other hand, cell state changes may be reversible: either stochastically, as in some embryonic stem cell cultures (Singer et al. 2014) or following chemical perturbations (Gupta et al. 2011), somatic cell nuclear transfer (Gurdon 1962), or reprogramming factors (Takahashi and Yamanaka 2006). Collectively, these processes exemplify transcriptional recombination. Branch-recombinant processes can also be observed at the onset of disease, for example, in dedifferentiation of cancer cells (Gupta et al. 2011; Friedmann-Morvinski and Verma 2014), or during host-pathogen interactions, when the pathogen seeks to supress the host's immune response and rewire metabolism for its own survival (Jenner and Young 2005; Lewis et al. 2015). Therefore, the ability to automatically quantify branch-recombinant processes, and thus identify the regulators of cell state transitions, has diverse applications in both fundamental and translational research.

The properties GPs make them an ideal framework for investigating these decision-making processes. However, whilst previous Gaussian process (GP)-based approaches to branching exist (Yang et al. 2016), they are applicable only for the two-branch case. Here we have developed branch-recombinant Gaussian processes (B-RGPs), a comprehensive statistical framework for the inference of branching and recombination, which can cope with an arbitrary number of bifurcations, and can incorporate a diverse range of dynamic behaviour. The ability to quantify more than two branching events simultaneously has relevance in biological systems. For instance, during gastrulation, three germ layers and PGCs are all specified within a short time period, which could be regarded as branching of pluripotent epiblast into four cell fates. Full characterisation of developmental dynamics from totipotent zygote to terminally differentiated cell types, incorporating all possible bifurcations, would require flexible approaches that can assign novel cells types to appropriate branches. Since our approach can cope with missing branch labels it represents an ideal framework for doing so, particularly if investigating understudied systems or species.

In general, we have demonstrated that B-RGPs are superior to standard GP regression for branching systems. The quantification of transcriptional branching and recombination during *Arabidopsis thaliana* response to a bacterial pathogen demonstrated its usefulness on genome-scale applications, and enabled us to separate out genes involved in the general immune response from those targeted by pathogen effectors and associated with the onset of disease. Furthermore, the precise computation of branch times allowed enriched identification of direct and indirect targets of key transcription factors involved in *Arabidopsis* response to bacterial infection.

Next, we used B-RGPs to investigate the dynamics of cell fate decision making in human primordial germ cells (hPGCs). By shortlisting genes that branched early between PGCs and soma, we were able to correctly identify key early regulators of PGC fate, most notably *SOX17*, a classic endoderm marker gene that was only recently shown to play a prominent role in human PGC specification (Irie et al. 2015; Tang et al. 2015). Furthermore, by jointly leveraging *in vivo*

and *in vitro* datasets, we were able to suggest the importance of correct suppression of *SOX2* and concomitant up-regulation of *SOX17* in hPGC fate. Interestingly, mouse PGCs do not require *Sox17* and instead express *Sox2*. It is possible that many more unexplored differences exist between mouse and human PGC transcriptomes (Tang et al. 2015). B-RGPs could prove useful in identifying such divergent regulators, whose roles could then be tested experimentally.

Our approach also identified a number of putative regulators for later stage PGC development, including *TNF* signalling. While the TNF pathway is primarily involved in pro-inflammatory apoptotic responses, emerging evidence points to its multiple morphogenetic roles (Mathew et al. 2009). Our analysis suggests that genes involved in TNF signalling differentially respond after PGC specification, but prior to expression of late germ cell markers, which could suggest TNF pathway's involvement in PGC migration. Indeed, TNFα is known to control cellular shape through modulating the cytoskeleton via Rho-family GTPases (Mathew et al. 2009), and the transcriptional signature of TNF signalling appears concomitant with SCF signalling, which has been implicated in PGC migration (De Felici 2013).

By looking at the earliest branching events, we identified putative regulators that may play a role prior to PGC specification, in the acquisition of PGC competence. Competence refers to a developmental potential to form a specific cell fate in response to particular signalling cues. Thus, WNT pathway has been implicated in endowing epiblast with competence to form PGCs in response to BMP2/4 (Ohinata et al. 2009; Kobayashi et al. 2017). However, both WNT and BMP signalling pathways are involved in specification of many somatic lineages in addition to PGCs (Mizutani and Bier 2008; Grigoryan et al. 2008). This could suggest the importance of cooperation with other signalling pathways, and therefore highlight the necessity of temporally resolved data. Here, we identified possible roles for FGFR signalling and its downstream branches (MEK/ERK, PI3K/AKT and JAK/STAT) in conferring human PGC competence. Importantly, these observations were validated experimentally using an *in vitro* system for derivation of PGC-like cells (PGCLCs) from hESCs. Thus, blocking FGFR or its downstream effectors (MEK, PI3K and JAK) by specific inhibitors enhanced the competency of hESCs to form PGCLCs. Thus our approach can inform genetic and signalling perturbations for cell fate decisions in a time-resolved manner.

Whilst we were able to correctly identify key events in the acquisition of competency and specification, key intermediate cell types were absent from both our *in vivo* and *in vitro* datasets. Notably, single cell RNA-seq for competent (4i) hESCs and PGCLCs were unavailable, although bulk RNA-seq measurements exist (Irie et al. 2015). A key future development will therefore aim to combine the use of single-cell and bulk RNA-seq data. Since bulk measurements represent population averages, B-RGPs would be ideally suited to this purpose, due to the ability of GPs to incorporate integral observations (Rasmussen and Williams 2006). A more informative approach would be single cell transcriptomic profiling of intermediate cell types, which would allow a much higher temporal resolution of the causal sequence of events leading to acquisition of competence and specification of hPGCs. The emergence of other *in vitro* models for human PGCLC derivation (Kobayashi et al. 2017) provide further opportunities for dissecting these causal regulations; by identifying differences and similarities in the dynamics of branching between different *in vitro* models, we could separate out the underlying biological mechanisms from culture-induced adaptations. Better time resolution would also be crucial for predicting direct targets of key PGC regulators, such as SOX17 and BLIMP1, which could be validated by

chromatin immunoprecipitation (ChIP) experiments.

In general, the use of B-RGPs to reconstruct cell fate dynamics using collections of single cell RNA-seq data can be combined with traditional gene-set enrichment analyses to provide a powerful framework for elucidating the mechanisms that underpin developmental processes. Whilst differential analysis alone may be used to identify putative regulators and perturbations, the temporal resolution provided by the B-RGPs significantly reduces the number of putative regulators, and concentrates on potentially key driving events, rather than enriched downstream effects (Supplementary Figure 12). Whilst the number of identified perturbations may still remain high, future studies may be able to leverage this type of approach alongside Bayesian optimisation strategies (Ghahramani 2015) to iteratively identify optimal culture conditions and mechanisms for deriving poorly characterised cell types.

Here we focused on transcriptional branching, but our framework is equally amenable to investigate the dynamics of other branching processes, including, for example, large scale epigenetic changes that occur during development, and the onset of disease. Indeed, recent studies have demonstrated the practicality of jointly sequencing the transcriptome and methylome from a single cell (Angermueller et al. 2016), which should provide opportunities to fully investigate the interlinked relationships of transcriptional programs and epigenetic state.

Finally, the transition from pluripotency to PGC competence and ultimately to PGCs can be reversed later in development upon germ cell tumour formation, exemplifying a recombination event. Seminoma and embryonal carcinoma are two types of germ cell tumours that share similarities with PGCs and ESC, respectively (Surani 2015), which can be distinguished using the markers identified from *in vitro* human PGCLC specification; seminomas and PGCs express CD38 and SOX17, while embryonal carcinomas and hESCs express SOX2 and CD30 (Irie et al. 2015). This underscores the importance of interrogating the transcriptional and epigenetic control of human germ cell fate and its specification from pluripotent progenitors. Ultimately, the use of B-RGPs on transcriptomics data from tumour cells could shed light on the sequence of events that lead to their formation, identify cancer markers, and guide therapeutic interventions.

Acknowledgements: We thank Ufuk Günesdogan and Naoko Irie for critical reading of the manuscript, Toshihiro Kobayashi for the provision of figures, and the Surani lab for useful insights. We thank Charles Bradshaw for help with high performance computing, and the Gurdon Institute core facilities for continued support. Finally, we thank Magnus Rattray and Jing Yang for useful insights and discussion of statistical branching models.

Funding: CAP is supported by Wellcome Trust grant (083089/Z/07/Z). AS is supported by a 4-year Wellcome Trust PhD in Scholarship and Cambridge International Trust Scholarship. MAS is supported by HFSP and a Wellcome Trust Senior Investigator Award. YH is supported by a studentship from the James Baird Fund, University of Cambridge. JR and LW are funded by the UK Medical Research Council (Grant Ref MC_U105260799). MG is supported by BBSRC grants BB/F005903/1 and BB/P002560/1. ZG acknowledges funding from the Alan Turing Institute, Google, Microsoft Research and EPSRC Grant EP/N014162/1.

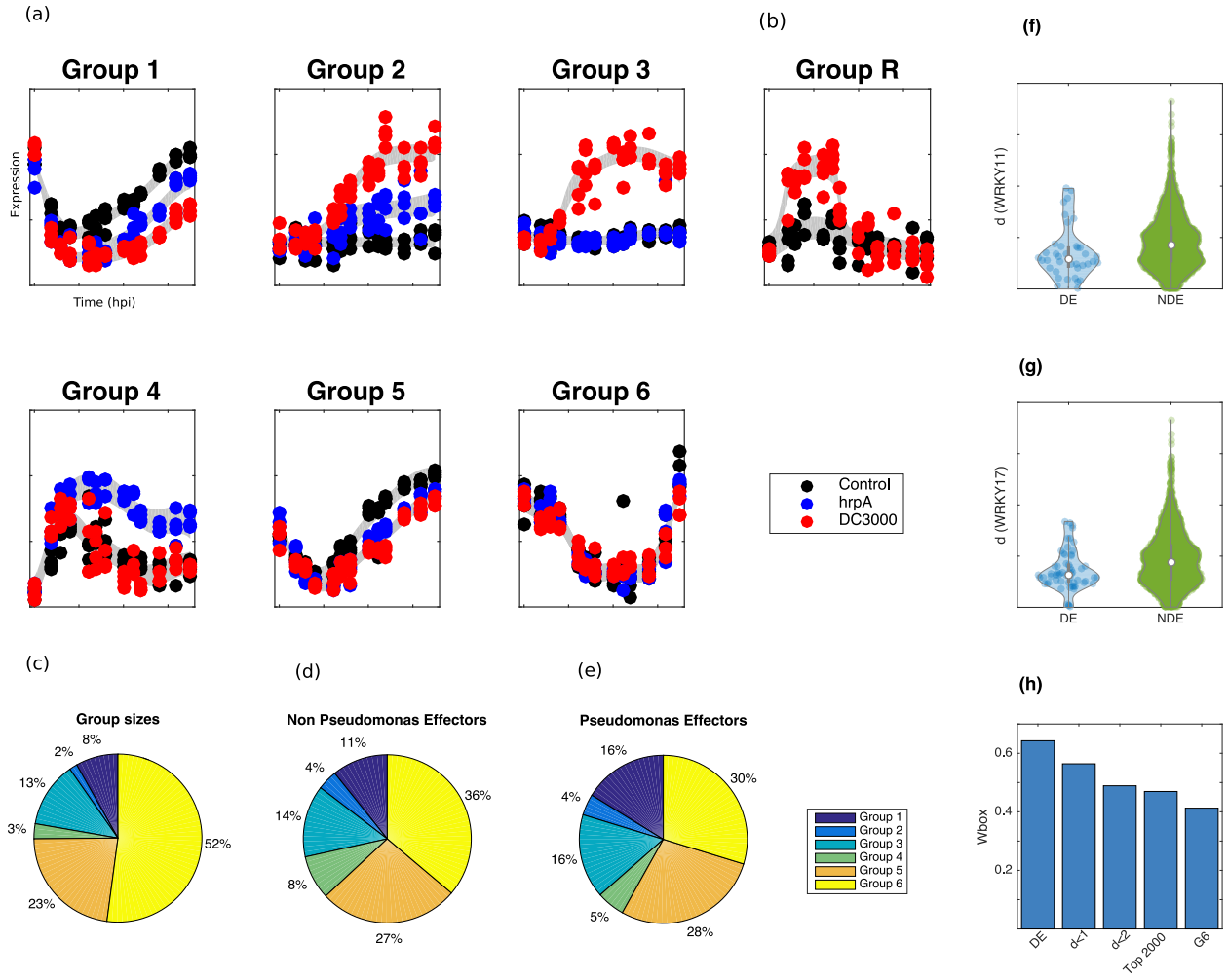


Figure 1: Branching processes were fitted to the three Arabidopsis time series, with hyper-parameters optimised to MAP values, and the BIC used to select optimal branching structure. (a) Example expression profile plots for each of the different classes of branching. **(b)** Example expression profile of a branch-recombinant structure within the dataset. **(c)** The prevalence of each of the six groups within the dataset, compared to the breakdown of non-Pseudomonas effector targets **(d)**, and Pseudomonas-effector targets show a clear enrichment of effector genes **(e)**. **(f, g)** The Euclidean distance of branching times of genes from that of WRKY11/17 is statistically lower in genes that are DE in WRKY11/17 knockouts versus those that are NDE, indicating that perturbation times are predictive of direct and indirect targets of WRKY11/17. **(h)** The prevalence of Wbox motifs decreases amongst sets of genes whose branch times are increasingly distant from WRKY11.

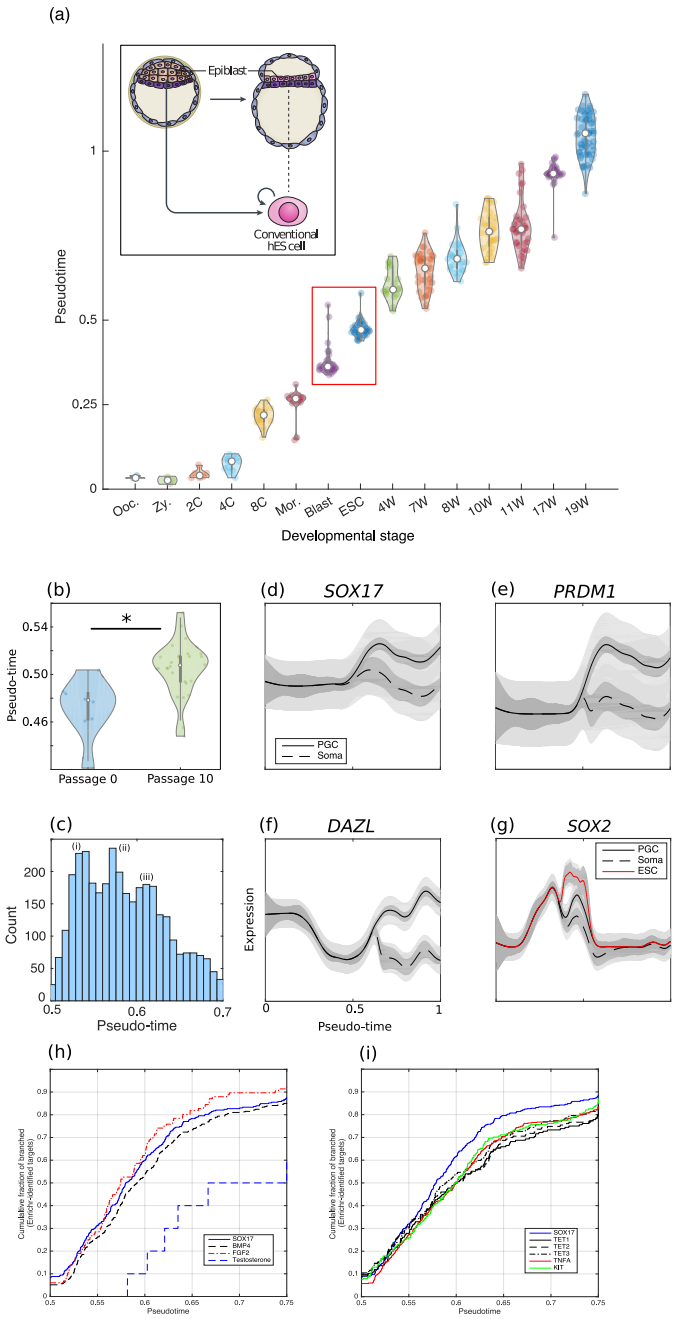


Figure 2: (a) Inferred pseudo-temporal ordering of individual cells over a two-component branching component correctly identifies the developmental ordering of blastocyst stage cells and ESCs, suggesting hESCs are developmentally more advanced than blastocyst stage. **(b)** ESCs at passage 10 appear to be developmentally more advanced than at earlier passages, having statistically later pseudotimes. **(c)** Histogram of branching time (soma versus PGCs) indicates a multimodal response. **(d, e, f)** Key PGC regulators *SOX17* and *PRDM1* branch early in the pseudotime series, prior to late PGC markers such as *DAZL*. **(g)** *SOX2*, a pluripotency and neuronal cell fate gene, appears to be branched between ESCs and the inferred *in vivo* dynamics. **(h)** Identified targets of *SOX17* and *BMP4* from perturbation studies in ESCs show early branching between PGCs and soma. **(i)** TNF response appears to be concomitant with genes associated with epigenetic reprogramming and the migratory phase of PGC development.

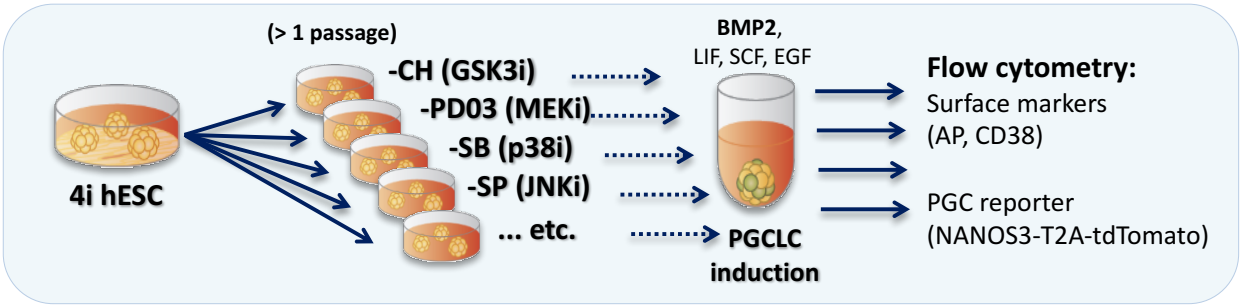


Figure 3. Scheme of the experimental workflow used to test the roles of individual components of the 4i hESC medium to sustain PGCLC competence. hESCs cultured in 4i medium were transferred to media lacking one of the kinase inhibitors for at least one passage. These cells were then subjected to standard PGCLC induction with BMP2 and supporting cytokines. PGCLC induction efficiency was assessed by flow cytometry as percentage of live cells expressing PGC markers AP, CD38 and NANOS3. Changes in PGCLC induction efficiency were thus used as proxy of changes in competence for PGC fate.

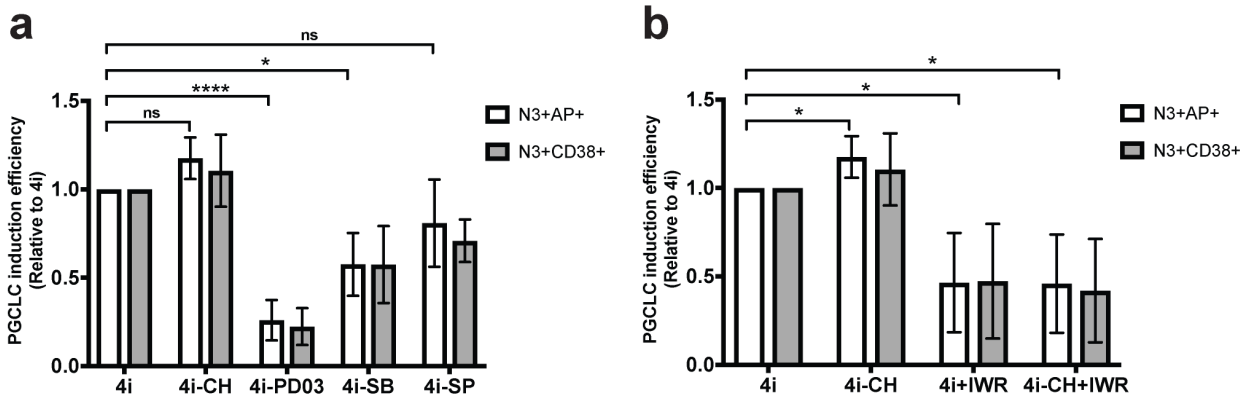


Figure 4. Quantification of PGCLC induction efficiency from hESCs grown in indicated conditions relative to 4i hESCs. PGCLCs induction efficiency was defined as the percentage of live NANOS3-tdTomato/AP (N3+AP+) or NANOS3-tdTomato/CD38-double positive cells. Data are shown as mean ± SD of 2 or 4 independent experiments. * $p \leq 0.05$, **** $P \leq 0.0001$, ns: not significant ($p > 0.05$), Holm-Sidak t-test (on relative frequency of N3+AP+ cells). **(a)** MEK and p38 inhibitors withdrawal from 4i medium decreases PGCLC competence of hESCs. **(b)** Inhibition of WNT signalling by small molecule inhibitor IWR reduces PGCLC competence.

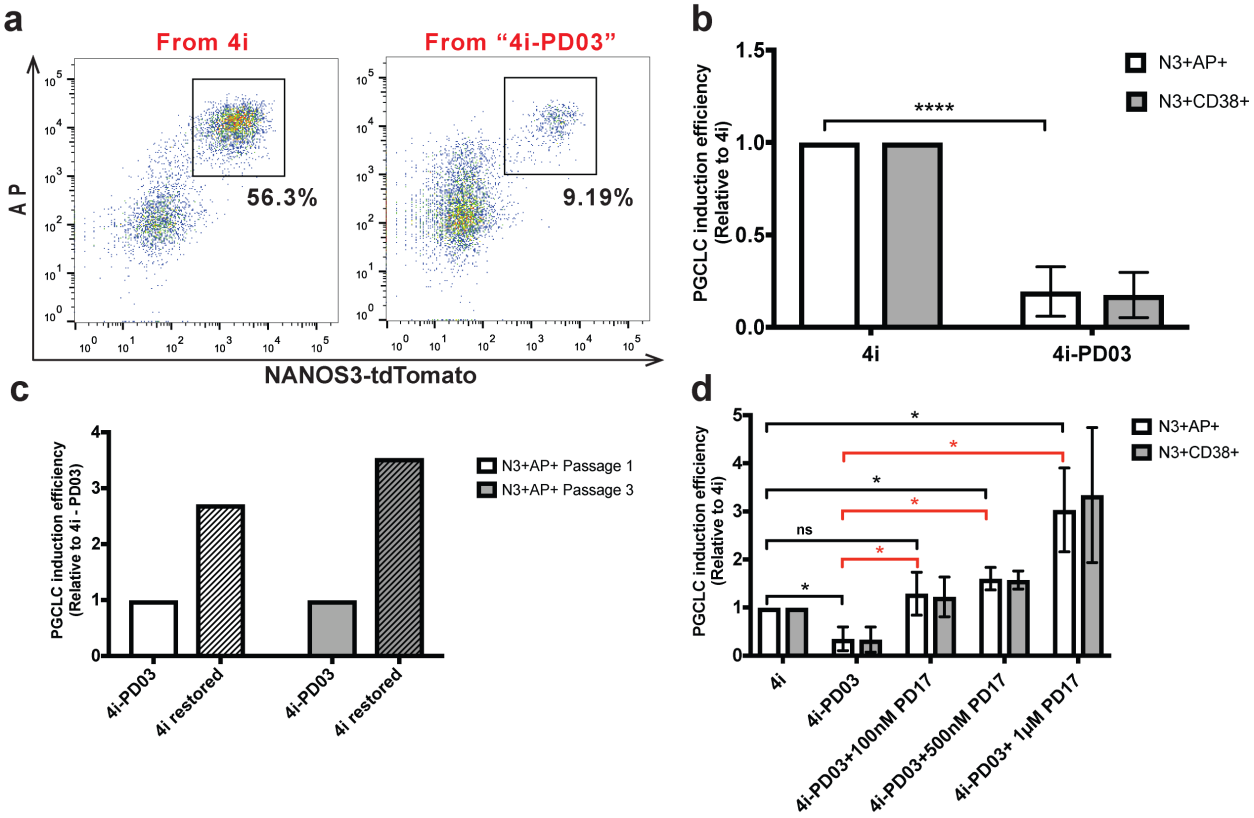


Figure 5. FGFR-MEK signalling is a negative regulator of human PGCLC competence. **(a)** MEK inhibitor withdrawal from 4i medium decreases PGCLC competence of hESCs. Representative flow cytometry plots of EBs derived from 4i and "4i-PD03" hESCs. **(b)** Quantification of PGCLC induction efficiency from "4i-PD03" relative to 4i hESCs. Data are shown as mean ± SD of 9 independent experiments. **** $p \leq 0.0001$, Holm-Sidak t-test (on relative frequency of N3+AP+ cells). **(c)** "4i-PD03" differentiation defect can be rescued by reintroducing PD03. hESCs grown in "4i-PD03" for >10 passages were transferred to complete 4i medium (with PD03) for 1 or 3 passages and subjected to PGCLC induction. Induction efficiency is shown relative to "4i-PD03"-derived PGCLCs; n=1. **(d)** hESCs grown with FGFR inhibitor PD17 show enhanced PGCLC competence. Data are shown as mean ± SD of 2 or 3 independent experiments. * $p \leq 0.05$, ns: not significant ($p > 0.05$), Holm-Sidak t-test (on relative frequency of N3+AP+ cells). Red lines and asterisks refer to comparison of "4i-PD03" to other conditions.

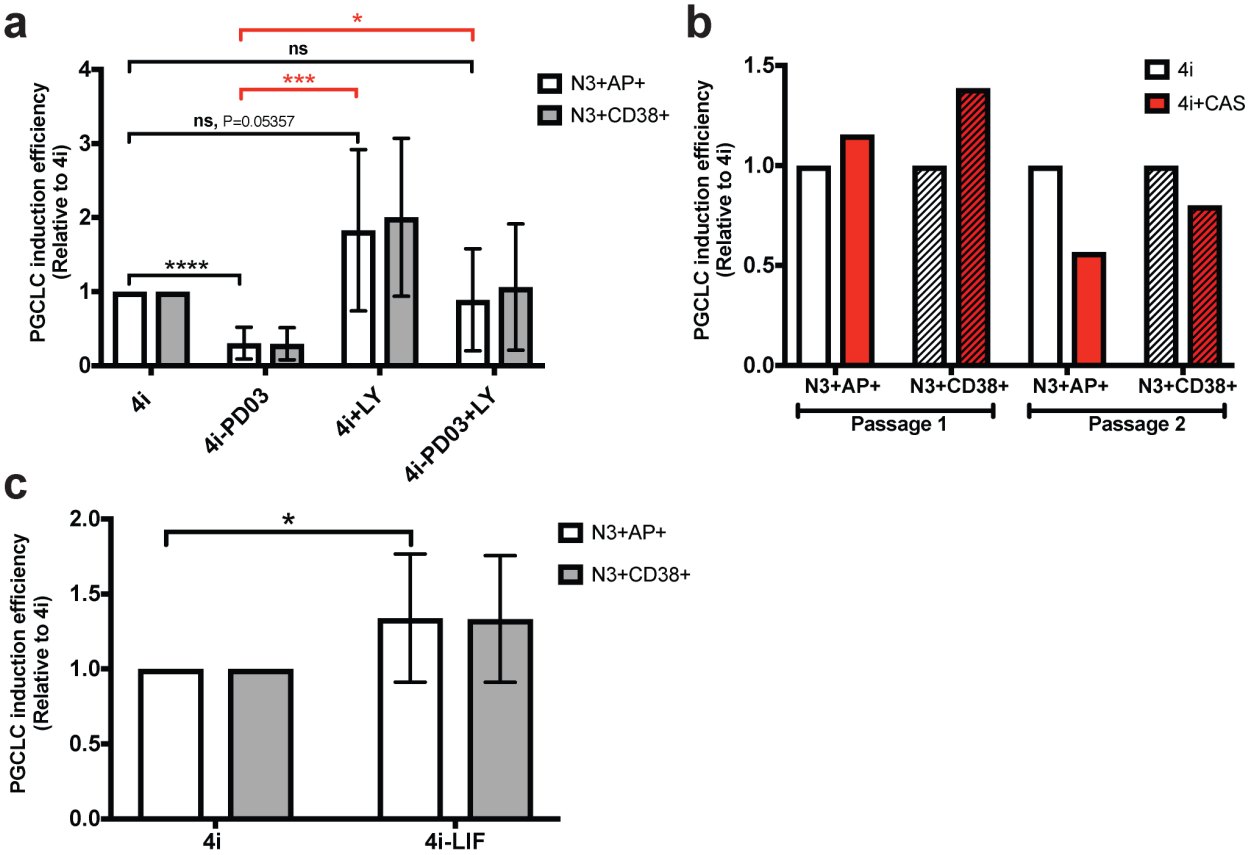


Figure 6. PI3K and LIF/JAK/STAT pathways negatively regulate PGCLC competence. **(a)** PI3K inhibition promotes PGCLC competence. Quantification of PGCLC induction efficiency from cells grown in indicated conditions relative to 4i hESCs. Data are shown as mean ± SD of 10 independent experiments. * $p \leq 0.05$, *** $p \leq 0.001$, **** $p \leq 0.0001$, ns: not significant ($p > 0.05$), Holm-Sidak t-test (on relative frequency of N3+AP+ cells). Red lines and asterisks refer to comparison of "4i-PD03" to other conditions. **(b)** Short exposure (1 passage) of hESCs to JAK inhibitor CAS increases PGCLC competence to form mature CD38+ PGCLCs. Quantification of PGCLC induction efficiency from cells grown in 4i+CAS relative to 4i hESCs. Longer culture with CAS changed colony morphology and decreased competence. **(c)** LIF withdrawal from 4i hESCs promoted PGCLC competence, although the magnitude of the effect was mild. Data are shown as mean ± SD of 10 independent experiments. * $p \leq 0.05$, Holm-Sidak t-test (on relative frequency of N3+AP+ cells).

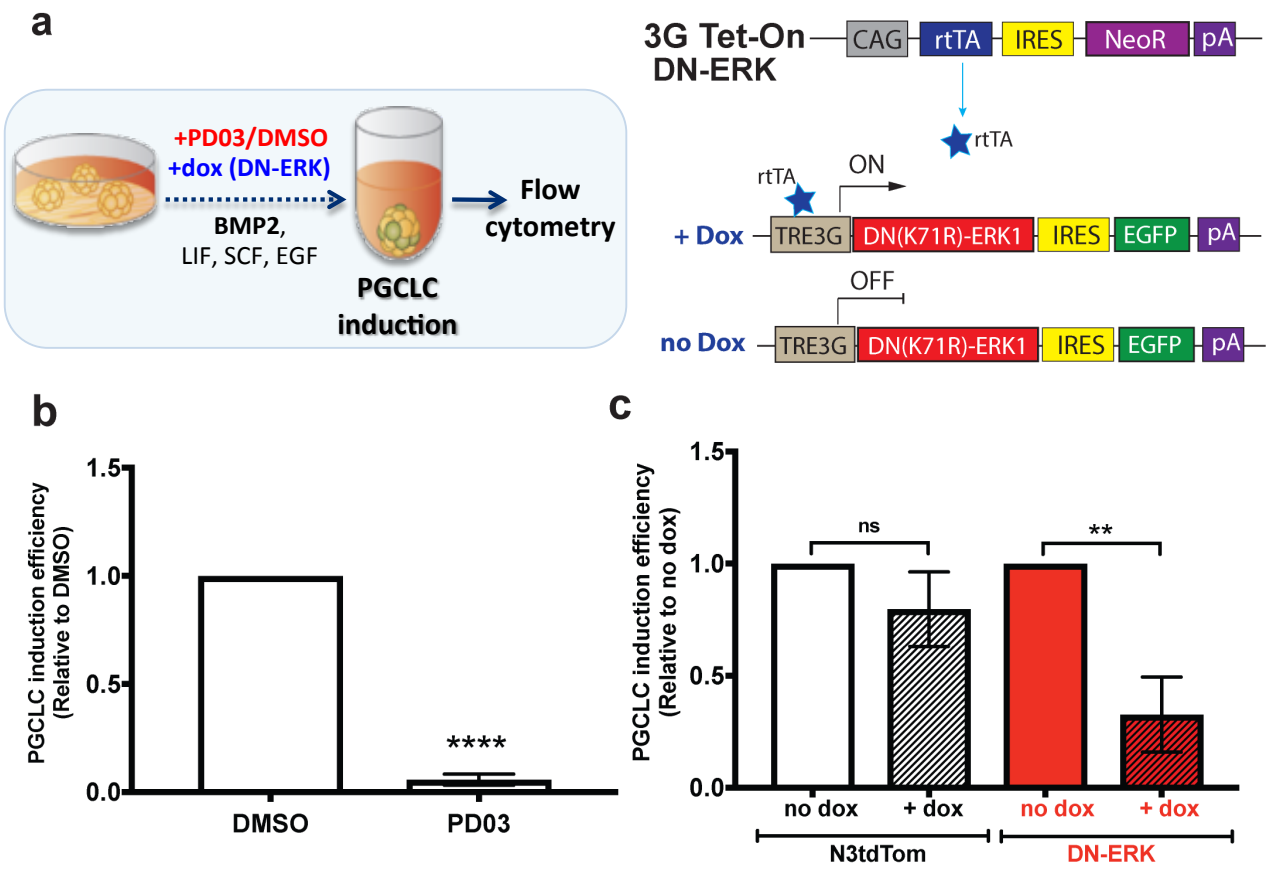


Figure 7. MEK-ERK inhibition during PGCLC induction interferes with PGCLC specification. **(a)** Workflow of the experiments and the scheme of the construct for DN-ERK transgene expression. Either PD03 or DMSO (vehicle) was added to the differentiation medium. Alternatively, to induce DN-ERK, PGCLC medium was supplemented with doxycycline (dox). **(b)** Supplementation of the differentiation medium with PD03 drastically decreases PGCLC specification efficiency. Data (relative frequency of live N3+AP+ cells) are shown as mean ± SD of 4 independent experiments. **** $P \leq 0.0001$, Holm-Sidak t-test. **(c)** DN-ERK expression during PGCLC differentiation reduces the efficiency of PGCLC specification. Data (relative frequency of live N3+AP+ cells) are shown as mean ± SD of 3 independent experiments using 2 DN-ERK-expressing clones. ** $p \leq 0.01$, Holm-Sidak t-test.

REFERENCES

Aksoy, I., R. Jauch, J. Chen, M. Dyla, U. Divakar, G. K. Bogu, R. Teo, C. K. Leng Ng, W. Herath, S. Lili, A. P. Hutchins, P. Robson, P. R. Kolatkar, and L. W. Stanton. 2013. 'Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm', *EMBO J*, 32: 938-53.

Angermueller, C., S.J. Clark, H.J. Lee, I.C. Macaulay, M.J. Teng, T.X. Hu, F. Krueger, S.A. Smallwood, C.P. Ponting, T. Voet, G. Kelsey, O. Stegle, and W. Reik. 2016. 'Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity', *Nature Methods*, 13: 229-32.

Aramaki, S., K. Hayashi, K. Kurimoto, H. Ohta, Y. Yabuta, H. Iwanari, Y. Mochizuki, T. Hamakubo, Y. Kato, K. Shirahige, and M. Saitou. 2013. 'A mesodermal factor, T, specifies mouse germ cell fate by directly activating germline determinants', *Dev Cell*, 27: 516-29.

Bailey, T. L. 2011. 'DREME: motif discovery in transcription factor ChIP-seq data', *Bioinformatics*, 27: 1653-9.

Borensztein, M., L. Syx, K. Ancelin, P. Diabangouaya, C. Picard, T. Liu, J. B. Liang, I. Vassilev, R. Galupa, N. Servant, E. Barillot, A. Surani, C. J. Chen, and E. Heard. 2017. 'Xist-dependent imprinted X inactivation and the early developmental consequences of its failure', *Nat Struct Mol Biol*, 24: 226-33.

Breeze, E., E. Harrison, S. McHattie, L. Hughes, R. Hickman, C. Hill, S. Kiddle, Y. S. Kim, C. A. Penfold, D. Jenkins, C. Zhang, K. Morris, C. Jenner, S. Jackson, B. Thomas, A. Tabrett, R. Legaie, J. D. Moore, D. L. Wild, S. Ott, D. Rand, J. Beynon, K. Denby, A. Mead, and V. Buchanan-Wollaston. 2011. 'High-resolution temporal profiling of transcripts during Arabidopsis leaf senescence reveals a distinct chronology of processes and regulation', *Plant Cell*, 23: 873-94.

Buettner, F., K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle. 2015. 'Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells', *Nat Biotechnol*, 33: 155-60.

Campbell, K., and C. Yau. 2015. 'Bayesian Gaussian Process Latent Variable Models for pseudotime inference in single-cell RNA-seq data', *bioRxiv*.

Campolo, F., M. Gori, R. Favaro, S. Nicolis, M. Pellegrini, F. Botti, P. Rossi, E. A. Jannini, and S. Dolci. 2013. 'Essential role of Sox2 for the establishment and maintenance of the germ cell line', *Stem Cells*, 31: 1408-21.

Carter, A. C., B. N. Davis-Dusenbery, K. Koszka, J. K. Ichida, and K. Eggan. 2014. 'Nanog-independent reprogramming to iPSCs with canonical factors', *Stem Cell Reports*, 2: 119-26.

Chambers, I., J. Silva, D. Colby, J. Nichols, B. Nijmeijer, M. Robertson, J. Vrana, K. Jones, L. Grotewold, and A. Smith. 2007. 'Nanog safeguards pluripotency and mediates germline development', *Nature*, 450: 1230-4.

Chen, B., M. E. Dodge, W. Tang, J. Lu, Z. Ma, C. W. Fan, S. Wei, W. Hao, J. Kilgore, N. S. Williams, M. G. Roth, J. F. Amatruda, C. Chen, and L. Lum. 2009. 'Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer', *Nat Chem Biol*, 5: 100-7.

Chen, E.Y., C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, N.R. Clark, and A. Ma'ayan. 2013. 'Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool', *BMC Bioinformatics*, 14: 128.

Ciolkowski, I., D. Wanke, R.P. Birkenbihl, and I.E. Somssich. 2008. 'Studies on DNA-binding selectivity of WRKY transcription factors lend structural clues into WRKY-domain function', *Plant Molecular Biology*, 68: 81-92.

De Felici, M. 2013. *Origin, Migration, and Proliferation of Human Primordial Germ Cells*.

Fong, H., K.A. Hohenstein, and P.J. Donovan. 2008. 'Regulation of self-renewal and pluripotency by Sox2 in human embryonic stem cells', *Stem Cells*, 26: 1931-38.

Friedmann-Morvinski, D., and I. M. Verma. 2014. 'Dedifferentiation and reprogramming: origins of cancer stem cells', *EMBO Rep*, 15: 244-53.

Gafni, O., L. Weinberger, A. A. Mansour, Y. S. Manor, E. Chomsky, D. Ben-Yosef, Y. Kalma, S. Viukov, I. Maza, A. Zviran, Y. Rais, Z. Shipony, Z. Mukamel, V. Krupalnik, M. Zerbib, S. Geula, I. Caspi, D. Schneir, T. Shwartz, S. Gilad, D. Amann-Zalcenstein, S. Benjamin, I. Amit, A. Tanay, R. Massarwa, N. Novershtern, and J. H. Hanna. 2013. 'Derivation of novel human ground state naive pluripotent stem cells', *Nature*, 504: 282-6.

Ghahramani, Z. 2015. 'Probabilistic machine learning and artificial intelligence', *Nature*, 521: 452-9.

Gkountela, S., K. X. Zhang, T. A. Shafiq, W. W. Liao, J. Hargan-Calvopina, P. Y. Chen, and A. T. Clark. 2015. 'DNA Demethylation Dynamics in the Human Prenatal Germline', *Cell*, 161: 1425-36.

Grant, C. E., T. L. Bailey, and W. S. Noble. 2011. 'FIMO: scanning for occurrences of a given motif', *Bioinformatics*, 27: 1017-8.

Grigoryan, T., P. Wend, A. Klaus, and W. Birchmeier. 2008. 'Deciphering the function of canonical Wnt signals in development and disease: conditional loss- and gain-of-function mutations of beta-catenin in mice', *Genes Dev*, 22: 2308-41.

Guo, F., L. Yan, H. Guo, L. Li, B. Hu, Y. Zhao, J. Yong, Y. Hu, X. Wang, Y. Wei, W. Wang, R. Li, J. Yan, X. Zhi, Y. Zhang, H. Jin, W. Zhang, Y. Hou, P. Zhu, J. Li, L. Zhang, S. Liu, Y. Ren, X. Zhu, L. Wen, Y. Q. Gao, F. Tang, and J. Qiao. 2015. 'The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells', *Cell*, 161: 1437-52.

Gupta, P. B., C. M. Fillmore, G. Jiang, S. D. Shapira, K. Tao, C. Kuperwasser, and E. S. Lander. 2011. 'Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells', *Cell*, 146: 633-44.

Gupta, S., J.A. Stamatoyannopoulos, T.L. Bailey, and W.S. Noble. 2007. 'Quantifying similarity between motifs', *Genome Biology*, 8: R24.

Gurdon, J. B. 1962. 'The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles', *Development*, 10: 622–40.

Hens, J., P. Dann, M. Hiremath, T. C. Pan, L. Chodosh, and J. Wysolmerski. 2009. 'Analysis of gene expression in PTHrP-/- mammary buds supports a role for BMP signaling and MMP2 in the initiation of ductal morphogenesis', *Dev Dyn*, 238: 2713-24.

Hensman, J., N.D. Lawrence, and M. Rattray. 2013. 'Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters', *BMC Bioinformatics*, 14: 252.

Hoyer, P. E., A. G. Byskov, and K. Mollgard. 2005. 'Stem cell factor and c-Kit in human primordial germ cells and fetal ovaries', *Mol Cell Endocrinol*, 234: 1-10.

Huang, Y., J. K. Kim, D. V. Do, C. Lee, C. A. Penfold, J. J. Zylicz, J. C. Marioni, J. A. Hackett, and M. A. Surani. 2017. 'Stella modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition', *Elife*, 6.

Irie, N., W. W. Tang, and M. Azim Surani. 2014. 'Germ cell specification and pluripotency in mammals: a perspective from early embryogenesis', *Reprod Med Biol*, 13: 203-15.

Irie, N., L. Weinberger, W. W. Tang, T. Kobayashi, S. Viukov, Y. S. Manor, S. Dietmann, J. H. Hanna, and M. A. Surani. 2015. 'SOX17 is a critical specifier of human primordial germ cell fate', *Cell*, 160: 253-68.

Ito, S., A. C. D'Alessio, O. V. Taranova, K. Hong, L. C. Sowers, and Y. Zhang. 2010. 'Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification', *Nature*, 466: 1129-33.

Jenner, R. G., and R. A. Young. 2005. 'Insights into host responses against pathogens from transcriptional profiling', *Nature Reviews Microbiology*, 3: 281-94.

Journot-Catalino, N., I. E. Somssich, D. Roby, and T. Kroj. 2006. 'The transcription factors WRKY11 and WRKY17 act as negative regulators of basal resistance in Arabidopsis thaliana', *Plant Cell*, 18: 3289-302.

Kalaitzis, A.A., and N.D. Lawrence. 2011. 'A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression', *BMC Bioinformatics*, 12: 180-93.

Kanai-Azuma, M., Y. Kanai, J.M. Gad, Y. Tajima, C. Taya, M. Kurohmaru, Y. Sanai, H. Yonekawa, K. Yazaki, P.P.L. Tam, and Y. Hayashi. 2002. 'Depletion of definitive gut endoderm in Sox17-null mutant mice', *Development*, 129: 2367-79.

Kawase, E., H. Yamamoto, K. Hashimoto, and N. Nakatsuji. 1994. 'Tumor necrosis factor-α (TNF-α) stimulates proliferation of mouse primordial germ cells in culture', *Developmental biology*, 161: 91-95.

Kempkensteffen, C., T. Jager, J. Bub, S. Weikert, S. Hinz, F. Christoph, H. Krause, M. Schostak, K. Miller, and M. Schrader. 2007. 'The equilibrium of XIAP and Smac/DIABLO expression is gradually deranged during the development and progression of testicular germ cell tumours', *Int J Androl*, 30: 476-83.

Kobayashi, T., H. Zhang, W.W.C. Tang, N. Irie, S. Withey, Klisch. Doris., A. Sybirna, S. Dietmann, D.A. Contreras, R. Webb, C. Allegrucci, R. Alberio, and M.A. Surani. 2017. 'Principles of early human development and germ cell program from conserved model systems', *Nature*.

Kuleshov, M. V., M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan. 2016. 'Enrichr: a comprehensive gene set enrichment analysis web server 2016 update', *Nucleic Acids Res*, 44: W90-7.

Lanner, F., and J. Rossant. 2010. 'The role of FGF/Erk signaling in pluripotent cells', *Development*, 137: 3351-60.

Lawrence, N.D. 2003. "Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data." In *Advances in Neural Information Processing Systems*, 5.

———. 2005. 'Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models', *Journal of Machine Learning Research*, 6: 1783–816.

Lawson, K.A., N.R. Dunn, B.A.J. Roelen, L.M. Zeinstra, A.M. Davis, C.V.E. Wright, J.P.W.F.M. Korving, and B.L.M. Hogan. 1999. 'Bmp4 is required for the generation of primordial germ cells in the mouse embryo', *GENES & DEVELOPMENT*, 13: 424–36.

Leitch, H. G., W. W. Tang, and M. A. Surani. 2013. 'Primordial germ-cell development and epigenetic reprogramming in mammals', *Curr Top Dev Biol*, 104: 149-87.

Lewis, L. A., K. Polanski, M. de Torres-Zabala, S. Jayaraman, L. Bowden, J. Moore, C. A. Penfold, D. J. Jenkins, C. Hill, L. Baxter, S. Kulasekaran, W. Truman, G. Littlejohn, J. Prusinska, A. Mead, J. Steinbrenner, R. Hickman, D. Rand, D. L. Wild, S. Ott, V. Buchanan-Wollaston, N. Smirnoff, J. Beynon, K. Denby, and M. Grant. 2015. 'Transcriptional Dynamics Driving MAMP-Triggered Immunity and Pathogen Effector-Mediated Immunosuppression in Arabidopsis Leaves Following Infection with Pseudomonas syringae pv tomato DC3000', *Plant Cell*, 27: 3038-64.

Lin, I. Y., F. L. Chiu, C. H. Yeang, H. F. Chen, C. Y. Chuang, S. Y. Yang, P. S. Hou, N. Sintupisut, H. N. Ho, H. C. Kuo, and K. I. Lin. 2014. 'Suppression of the SOX2 neural effector gene by PRDM1 promotes human germ cell fate in embryonic stem cells', *Stem Cell Reports*, 2: 189-204.

Magnusdottir, E., S. Dietmann, K. Murakami, U. Gunesdogan, F. Tang, S. Bao, E. Diamanti, K. Lao, B. Gottgens, and M. Azim Surani. 2013. 'A tripartite transcription factor network regulates primordial germ cell specification in mice', *Nat Cell Biol*, 15: 905-15.

Makoolati, Z., M. Movahedin, and M. Forouzandeh-Moghadam. 2016. 'Proliferation in culture of primordial germ cells derived from embryonic stem cell: induction by retinoic acid', *Biosci Rep*, 36.

Masui, S., Y. Nakatake, Y. Toyooka, D. Shimosato, R. Yagi, K. Takahashi, H. Okochi, A. Okuda, R. Matoba, A. A. Sharov, M. S. Ko, and H. Niwa. 2007. 'Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells', *Nat Cell Biol*, 9: 625-35.

Mathew, S. J., D. Haubert, M. Kronke, and M. Leptin. 2009. 'Looking beyond death: a morphogenetic role for the TNF signalling pathway', *J Cell Sci*, 122: 1939-46.

Mizutani, C. M., and E. Bier. 2008. 'EvoD/Vo: the origins of BMP signalling in the neuroectoderm', *Nat Rev Genet*, 9: 663-77.

Mollgard, K., A. Jespersen, M. C. Lutterodt, C. Yding Andersen, P. E. Hoyer, and A. G. Byskov. 2010. 'Human primordial germ cells migrate along nerve fibers and Schwann cells from the dorsal hind gut mesentery to the gonadal ridge', *Mol Hum Reprod*, 16: 621-31.

Mostert, M.C. , A.J.M.H. Verkerk, M. van de Pol, J. Heighway, P. Marynen, C. Rosenberg, A.G. van Kessel, J. van Echten, B. de Jong, J.W. Oosterhuis, and L.H.J. Looijenga. 1998. 'Identification of the critical region of 12p over-representation in testicular germ cell tumors of adolescents and adults', *Oncogene*, 16: 2617-27.

Mukhtar, M.S., A. Carvunis, M. Dreze, P. Epple, J. Steinbrenner, J. Moore, M. Tasan, M. Galli, T. Hao, M.T. Nishimura, S.J. Pevzner, S.E. Donovan, L. Ghamsari, B. Santhanam, V. Romero, M.M. Poulin, F. Gebreab, B.J. Gutierrez, S. Tam, D. Monachello, M. Boxem, C.J. Harbort, N. McDonald, L. Gai, H. Chen, Y. He, European Union Effectoromics Consortium, J. Vandenhaute, F.P. Roth, D.E. Hill, J.R. Ecker, M. Vidal, J. Beynon, P. Braun, and J.L. Dangl. 2011. 'Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network', *Science*, 333: 596–601.

Murakami, K., U. Gunesdogan, J. J. Zylicz, W. W. Tang, R. Sengupta, T. Kobayashi, S. Kim, R. Butler, S. Dietmann, and M. A. Surani. 2016. 'NANOG alone induces germ cells in primed epiblast in vitro by activation of enhancers', *Nature*, 529: 403-7.

Nakaki, F., K. Hayashi, H. Ohta, K. Kurimoto, Y. Yabuta, and M. Saitou. 2013. 'Induction of mouse germ-cell fate by transcription factors in vitro', *Nature*, 501: 222-6.

Ohinata, Y., H. Ohta, M. Shigeta, K. Yamanaka, T. Wakayama, and M. Saitou. 2009. 'A signaling principle for the specification of the germ cell lineage in mice', *Cell*, 137: 571-84.

Onishi, K., and P. W. Zandstra. 2015. 'LIF signaling in stem cells and development', *Development*, 142: 2230-6.

Petropoulos, S., D. Edsgard, B. Reinius, Q. Deng, S. P. Panula, S. Codeluppi, A. Plaza Reyes, S. Linnarsson, R. Sandberg, and F. Lanner. 2016. 'Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos', *Cell*, 165: 1012-26.

Rasmussen, C. E., and C. K. Williams. 2006. *Gaussian processes for machine learning* (MIT Press).

Reid, J.E., and L. Wernisch. 2016. 'Pseudotime estimation: deconfounding single cell time series', *Bioinformatics*, 32: 2973-80.

Sandberg, A.A., A.M. Meloni, and R.F. Suijkerbuijk. 1996. 'Reviews of Chromosome Studies in Urological Tumors. III. Cytogenetics and Genes in Testicular Tumors', *The Journal of Urology*, 55: 1531-56.

Sandlow, J.I., H. Feng, M.B. Cohen, and A. Sandra. 1996. 'Expression of c-KIT and Its Ligand, Stem Cell Factor, in Normal and Subfertile Human Testicular Tissue', *Journal of Andrology*, 17: 403-08.

Sasaki, K., S. Yokobayashi, T. Nakamura, I. Okamoto, Y. Yabuta, K. Kurimoto, H. Ohta, Y. Moritoki, C. Iwatani, H. Tsuchiya, S. Nakamura, K. Sekiguchi, T. Sakuma, T. Yamamoto, T. Mori, K. Woltjen, M. Nakagawa, T. Yamamoto, K. Takahashi, S. Yamanaka, and M. Saitou. 2015. 'Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells', *Cell Stem Cell*, 17: 178-94.

Schlessinger, J. 2004. 'Common and distinct elements in cellular signaling via EGF and FGF receptors', *Science*, 306: 1506-7.

Seguin, C. A., J. S. Draper, A. Nagy, and J. Rossant. 2008. 'Establishment of endoderm progenitors by SOX transcription factor expression in human embryonic stem cells', *Cell Stem Cell*, 3: 182-95.

Singer, Z. S., J. Yong, J. Tischler, J. A. Hackett, A. Altinok, M. A. Surani, L. Cai, and M. B. Elowitz. 2014. 'Dynamic heterogeneity and DNA methylation in embryonic stem cells', *Mol Cell*, 55: 319-31.

Stegle, O., K.J. Denby, E.J. Cooke, D.L. Wild, Z. Ghahramani, and K.M. Borgwardt. 2010. 'A Robust Bayesian Two-Sample Test for Detecting Intervals of Differential Gene Expression in Microarray Time Series', *JOURNAL OF COMPUTATIONAL BIOLOGY*, 17: 355–67.

Surani, M. A. 2015. 'Human Germline: A New Research Frontier', *Stem Cell Reports*, 4: 955-60.

Takahashi, K., and S. Yamanaka. 2006. 'Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors', *Cell*, 126: 663-76.

Tang, W. W., S. Dietmann, N. Irie, H. G. Leitch, V. I. Floros, C. R. Bradshaw, J. A. Hackett, P. F. Chinnery, and M. A. Surani. 2015. 'A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development', *Cell*, 161: 1453-67.

Tang, W. W., T. Kobayashi, N. Irie, S. Dietmann, and M. A. Surani. 2016. 'Specification and epigenetic programming of the human germ line', *Nat Rev Genet*, 17: 585-600.

Thomson, M., S. J. Liu, L. N. Zou, Z. Smith, A. Meissner, and S. Ramanathan. 2011. 'Pluripotency factors in embryonic stem cells regulate differentiation into germ layers', *Cell*, 145: 875-89.

Wagner, R., and R.R Myers. 1996. 'Schwann cells produce tumor necrosis factor alpha: expression in injured and non-injured nerves', *Neuroscience*, 73: 625-29.

Yamaguchi, S., K. Kurimoto, Y. Yabuta, H. Sasaki, N. Nakatsuji, M. Saitou, and T. Tada. 2009. 'Conditional knockdown of Nanog induces apoptotic cell death in mouse migrating primordial germ cells', *Development*, 136: 4011-20.

Yan, L., M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang. 2013. 'Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells', *Nat Struct Mol Biol*, 20: 1131-9.

Yang, J., C.A. Penfold, M.R. Grant, and M. Rattray. 2016. 'Inferring the perturbation time from biological time course data', *Bioinformatics*, 32: btw329.

Ying, Q. L., J. Wray, J. Nichols, L. Batlle-Morera, B. Doble, J. Woodgett, P. Cohen, and A. Smith. 2008. 'The ground state of embryonic stem cell self-renewal', *Nature*, 453: 519-23.

Zwiessele, M., and N.D. Lawrence. 2016. 'Topslam: Waddington Landscape Recovery for Single Cell Experiments', *bioRxiv*, 057778.

Supplementary materials to 'Nonparametric Bayesian inference of transcriptional branching and recombination identifies regulators of early human germ cell development'

Christopher A. Penfold[1,*,‡], Anastasiya Sybirna[1,2,3,*], John Reid[4], Yun Huang[1,3], Lorenz Wernisch[4], Zoubin Ghahramani[5], Murray Grant[6], and M. Azim Surani[1,2,3,§]

[1]Wellcome/CRUK UK Gurdon Institute, University of Cambridge
[2] Wellcome/MRC Cambridge Stem Cell Institute, University of Cambridge
[3]Physiology, Development and Neuroscience Department, University of Cambridge
[4]MRC Biostatistics Unit, University of Cambridge
[5]Department of Engineering, University of Cambridge
[6]School of Life Sciences, University of Warwick
[*]These authors contributed equally
[‡]c.penfold@gurdon.cam.ac.uk, [§]a.surani@gurdon.cam.ac.uk

1 BRANCH-RECOMBINANT GAUSSIAN PROCESSES (B-RGPS)

Within a Bayesian setting, Gaussian processes (GPs) can be used to represent prior distributions over smooth functions, providing a flexible framework for regression and classification with robust treatment of uncertainty (Rasmussen and Williams 2006). For regression, we typically have a set of observations, $\boldsymbol{y}$, assumed to be noisy instances of a continuous underlying function at input locations $\boldsymbol{t}$:

$$\boldsymbol{y} = f(\boldsymbol{t}) + \epsilon,$$

where ε represents Gaussian additive noise. In our applications, $\boldsymbol{y}$ will typically be used to indicate a vector of the observed expression level of a given gene at times, $\boldsymbol{t}$. We can assign the unknown function a GP prior, denoted $f(t) \sim \mathcal{GP}(\mu(t), \mathcal{C}(t, t'))$, and analytically evaluate the posterior distribution at a set of new input locations, $\boldsymbol{t}^*$. The marginal likelihood, too, may be analytically evaluated, making GPs a flexible and efficient framework for both prediction and model comparison. Previous GP-based approaches to branching have been outlined for the two-dataset case i.e., where there exists two biological processes following branching. These include the studies by Stegle et al. (2010), who developed a GP two-sample approach based on mixtures of GPs, and more recent studies by Yang et al. (2016), who demonstrate how an explicit two-branch process can be encoded within a joint GP model for the squared-exponential covariance function. To our knowledge, the generalisation of GPs to more than one branch has not currently been adequately addressed, with no treatment for recombinant processes.

A useful extension to the GP framework is the multiple output Hierarchical Gaussian Processes (HGP; Hensman, Lawrence, and Rattray (2013)), in which a basal process is defined by a zero-mean GP with covariance function $k_1(t, t')$, with a subsequent process having mean $f_1(t)$ and covariance function, $k_2(t, t')$:

$$f_1(t) \sim \mathcal{GP}\big(0, k_1(t, t')\big),$$

$$f_2(t) \sim \mathcal{GP}\big(f_1(t), k_2(t,t')\big).$$

Within this framework we assume noisy observations of the functions, $y_1 = f_1(t) + \epsilon$, and $y_2 = f_2(t) + \epsilon$, and may analytically evaluate the posterior distribution at a new set of input locations for prediction, or the marginal likelihood for model comparison. More general models, involving multiple layers of hierarchy can be built in this way (Hensman, Lawrence, and Rattray 2013).

A class of branching behaviour can naturally be encoded within this HGP framework, by assuming the basal (main branch) process is defined by a zero-mean GP with covariance function $k_{b_1}(t,t')$, with a subsequent process having mean $f_{b_1}(t)$ and an appropriate covariance function:

$$f_{b_1}(t) \sim \mathcal{GP}\left(0, k_{b_1}(t,t')\right),$$
$$f_{b_2}(t) \sim \mathcal{GP}\big(f_{b_1}(t), CP(K_0, k_{b_2}(t,t'))\big),$$

where $K_0 = K_0(t,t')$ denotes a zero-kernel, and $CP(K_1, k_2)$ denotes a change-point kernel (Lloyd et al. 2014):

$$CP\big(K_1, k_2(t,t')\big) = \sigma(t)k_1(t,t')\sigma(t') + \big(1-\sigma(t)\big)k_2(t,t')\big(1-\sigma(t')\big),$$

where $\big(1-\sigma(t)\big) = 1 + \tanh\left(\frac{t_b - t}{s}\right)/2$. Each datapoint is assigned a branch label, $z \in [1,2]$, and observations are *a priori* Gaussian distributed: $\boldsymbol{y}_{b_1}, \boldsymbol{y}_{b_2} | \, \boldsymbol{t}, \boldsymbol{z} \sim N(\boldsymbol{0}, K(\boldsymbol{t}, \boldsymbol{t}', \boldsymbol{z}, \boldsymbol{z}'))$, where:

$$K(t,t',z,z') = K_{b_1}(t,t') + CP\big(K_0, K_{b_2}(t,t')\big)\delta_{z,2}\delta_{z',2} + \beta\delta_{tz}\delta_{t'z'},$$

and the delta function $\delta_{z,2}\delta_{z',2}$ ensures the change-point kernel only operates over the second branch i.e., where the branch label $z = 2$ and $z' = 2$. Within this framework, we may again analytically evaluate a prediction at unknown input location, (t^*, z^*), as well as analytically evaluate the marginal likelihood, allowing us to compare the goodness of fit between different branching structures. We could allow further branches that independently diverge from the main branch, with an (n + 1)-component system defined by the following covariance function:

$$K(t,t',z,z') = K_{b_1}(t,t') + \textstyle\sum_{i=1}^{n} CP_{b_i}\big(K_0, K_{b_2}(t,t')\big)\delta_{z,i}\delta_{z',i} + \beta\delta_{tz}\delta_{t'z'}.$$

An example three-component system is illustrated in Supplementary Figure 1(c, d). Finally, each branch itself could give rise to further branching in a recurrent manner e.g., a basal (main branch) function from which a secondary branch diverges, with a third processes branching from the second and so forth. For an (n+1)-component recurrent system we have:

$$K(t,t',z,z') = K_{b_1}(t,t') + \textstyle\sum_{j=1}^{n}\big(\sum_{i=1}^{j} CP_{b_i}\big(K_0, K_{b_2}(t,t')\big)\big)\delta_{z,j}\delta_{z',j} + \beta\delta_{tz}\delta_{t'z'}.$$

Far more complex branching patterns can easily be built via the correct composition of independent and recurrent branching covariance functions. Note here, that the base-kernel and kernels associated with individual branches may, themselves, by arbitrarily complex, comprised of any linear combination of positive semidefinite kernels.

In Supplementary Figure 1(a) we indicate example behaviour for simple branching GPs. Here we include a heatmap representation of the covariance function evaluated at 1,000 uniformly incremented test locations (leftmost column), and two independently drawn samples from the prior distribution at those test locations (middle and rightmost columns).

1.1 RECOMBINANT GAUSSIAN PROCESSES

Recombination processes can be defined in a reciprocal fashion to branching processes. For example, we can describe a two-component system via the following composition of covariance functions:

$$K(t,t',z,z') = K_{b_1}(t,t') + CP_{r_2}\big(K_{b_2}(t,t'),K_0\big)\delta_{z,2}\delta_{z',2} + \beta\delta_{tz}\delta_{t'z'}$$

which encodes a main branch process, with a second, potentially correlated process that recombines after time r_1. Multiple processes can be allowed to independently recombine with the main branch, or recurrently recombine via a series of parental branches, analogously to branching GPs. Example recombinant GPs are shown in Supplementary Figure 1(b).

1.2 BRANCH-RECOMBINANT GAUSSIAN PROCESSES

Another important process exists where a statistical process transiently branches into two or more processes before recombining back into a single process. An example two-component system can be encoded by the following covariance function:

$$K(t,t',z,z') = K_{b_1}(t,t') + CP_{r_2}\big(CP_{b_2}\big(K_0,K_{b_2}(t,t')\big),K_0\big)\delta_{z,2}\delta_{z',2} + \beta\delta_{tz}\delta_{t'z'}$$

Again, more complex patterns, with arbitrary numbers of branches and recombination, can readily be built with GPs via the correct composition of covariance functions, with more complex examples shown in Supplementary Figures 1(c).

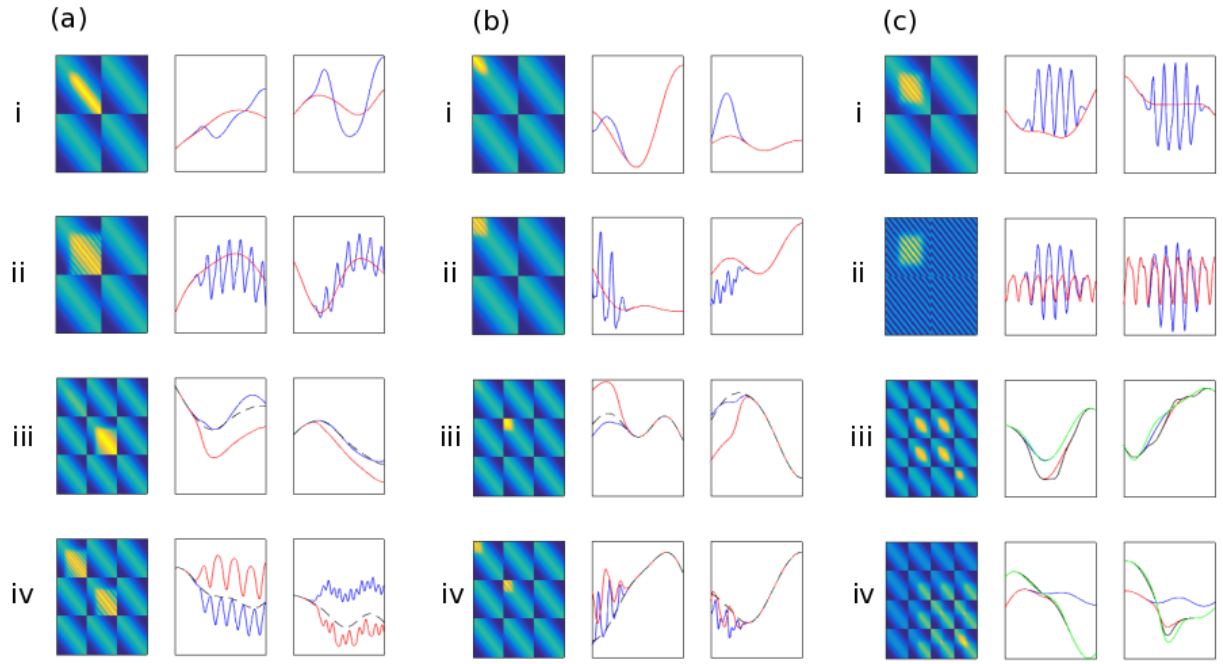
1.3 LIMITING BEHAVIOUR

Depending upon the branch time and other hyperparameters in the change-point kernel, the behaviour of BGPs can naturally tend towards either an independent GP or a HGP. Specifically, when $\sigma(t) \to 1$, as may be the case when $(t_b - t)/s$ is very large such as when a branch occurs much later than the last data point, then a BGP will behave as single joint GP with behaviour defined by the main branch kernel only. When $\sigma(t) \to 0$, as may be the case when $(t_b - t)/s$ has increasingly lower values, such as when branching occurs much earlier than the first observations, then the BGP will behave as a HGP. Similar limiting behaviour applies for recombination processes and for branching/recombinant processes.

1.4 CODE AND IMPLEMENTATION

B-RGPs have been implemented in MATLAB using the gpml toolbox (Rasmussen and Williams (2006); http://www.gaussianprocess.org/gpml/code/matlab/doc), allowing B-RGPs to be fitted to data with hyperparameters optimised using type II maximum likelihood (ML) estimates, *maximum a posteriori* (MAP) estimates, or else sampled using Markov chain Monte Carlo (MCMC). Code

for implementing B-RGPs and repeating experiments in the main text have been deposited at GitHub (https://github.com/cap76/BranchingGPs). An interactive Jupyter notebook is also available.



Supplementary Figure 1: (a) GP with branching structure. Here we indicate the prior covariance matrix evaluated at uniformly incremented input times (left) and two samples from the prior distribution (middle, right). (i) A smooth function branches from a smooth base function. (ii) A periodic function branches from a smooth function. (iii) Two smooth functions (solid) branch from a smooth function (dashed). (iv) Two periodic functions (solid) branch from a smooth function (dashed). **(b) GP with recombinant structure.** (i) A smooth function converges on another smooth function. (ii) A periodic function recombines with a smooth function. (iii) Two smooth functions recombine with a smooth function. (iv) Two periodic functions recombine with a smooth function. **(c) GP with branch-recombinant structure.** (i) A periodic function branches and then recombines with a smooth function. (ii) Two periodic functions branch and then recombine with one another. (iii) A set of four smooth functions branch and recombine at various time points. (iv) A set of two functions (black and green) branch and recombine with one another, whilst a third function (red) recombines with the first two functions, with a fourth function (blue) branching from the third function prior to its recombination with the first two.

2 BENCHMARKING OF B-RGPS

2.1 INFERENCE OF B-RGPS FROM LABELLED DATA

As a preliminary test of the B-RGP framework we fitted four simulated, labelled datasets, and evaluated the predictive accuracy over a range of test locations, comparing the accuracy to that achieved using independent Gaussian process regression over separate branches, or a joint Gaussian process applied to the union of data. In Experiment 1 we evaluated the ability to fit the following branch process:

$$f(t,z) = \begin{cases} 0, \text{ if } t \leq -\pi/2, \\ \cos(t), \text{ if } -\pi/2 < t \leq 0, z = 1, \\ 1, \text{ if } t > 0, z = 1, \\ -\cos(t), \text{ if } -\pi/2 < t \leq 0, z = 2, \\ -1, \text{ if } t > 0, z = 2. \end{cases}$$

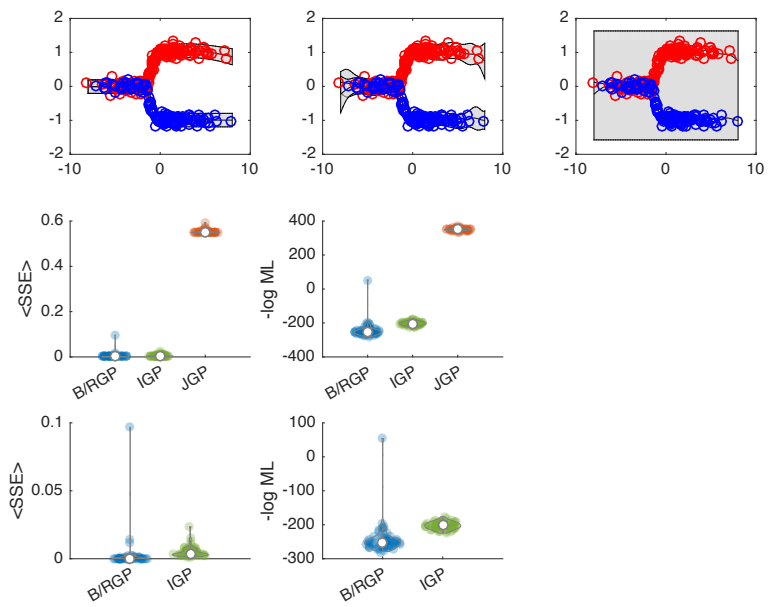
where z indicates the branch label. Three-hundred random input locations were sampled, $\boldsymbol{t} \sim \mathcal{N}(\mathbf{0}, 3\mathbb{I})$, with branch labels assigned with equal probability, $\boldsymbol{z}_i \in [1,2]$. Observations were generated as noisy instances, $y|t,z \sim \mathcal{N}(f(t,z), \sigma_n^2\mathbb{I})$, where $\sigma_n = 0.1$. A three-component branching process, comprised of a (latent) main process from which two observed branches diverge, was fitted to the simulated data, with hyperparameters optimised using type II maximum likelihood (ML). Here the base kernel for the main branch was assumed to be a squared-exponential, with the covariance function for the two branches also squared-exponentials. Branching time hyperparameters were tied i.e., $t_{b_1} = t_{b_2}$, with initial values set as $t_{b_1} = 4$, $\log s_{b_1,b_2} = 0.5$, $\sigma_n = 0.2$, and all other hyperparameters $\theta = [l_{b_0}, \sigma_{b_0}, l_{b_1}, \sigma_{b_1}, l_{b_2}, \sigma_{b_2}]$ initiated as $\theta_i \sim U(0.1,1)$. For comparison, independent Gaussian processes (IGP) were fitted to the two branches separately using squared-exponential covariance functions, with hyperparameters $\theta = [l_1, \sigma_1, l_2, \sigma_2]$ initiated as $\theta_i \sim U(0.1,1)$, and optimised using type II ML estimation. A joint GP (JGP) with squared-exponential covariance function was also fitted to the union of data, with hyperparameters $\theta = [l, \sigma]$ initiated and optimised as for the independent GPs. In Figure 3 (top) we indicate an example posterior GP fit to the data using a BGP, IGPs and JGP respectively. In Figure 3 (middle and bottom row) we indicate the mean sum of squared error (SSE) and negative log marginal likelihood over 50 randomly initiated runs. The B-RGP shows increased accuracy (reduced SSE) and decreased negative log marginal likelihood compared to the individual and joint GPs.

In Experiment 2 we assumed the following branch-recombinant process:

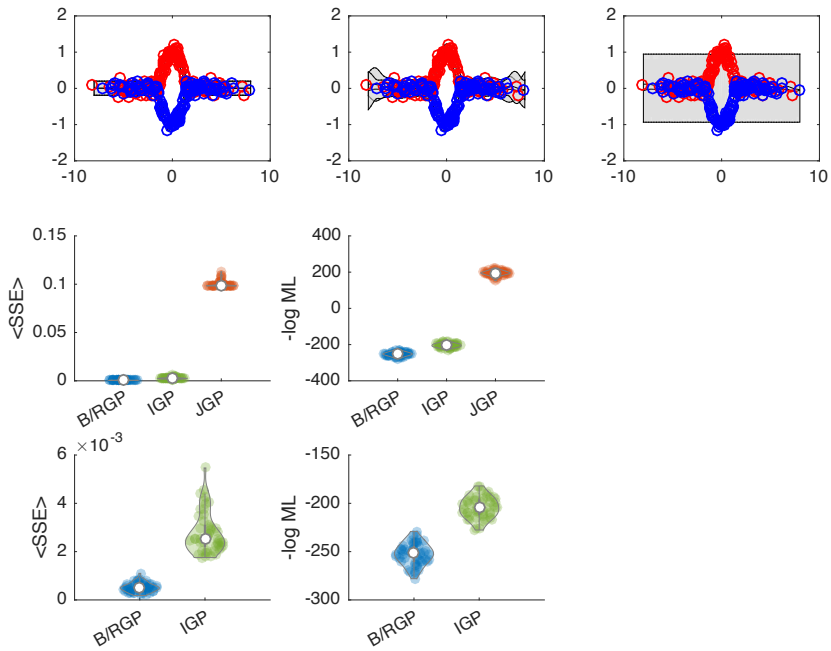
$$f(t,z) = \begin{cases} 0 \text{ if } |t| > \pi/2 \\ \cos(t) \text{ if } |t| \leq \pi/2, z = 1 \\ -\cos(t) \text{ if } |t| \leq \pi/2, z = 2 \end{cases}$$

where z indicates the branch label. Again, three-hundred randomly determined input locations were sampled $t \sim \mathcal{N}(0,3\mathbb{I})$, with branch assignment randomly allocated with equal probability, and

observation data generated as $y|t,z \sim \mathcal{N}(f(t,z), \sigma_n^2 \mathbb{I})$, where $\sigma_n = 0.1$. A three-component branch-recombinant GP, comprised of a (latent) main process from which two branches diverge and recombine, was fitted to the simulated data, with hyperparameters optimised using type II ML estimates. Here the base kernel for the main branch was assumed to be a squared-exponential, with the covariance function for the two branches also assumed to be squared-exponentials. Branching and recombination times for the two branches were tied i.e., $t_{b_1} = t_{b_2}$ and $t_{r_1} = t_{r_2}$, with initial values set as $t_{b_1} = -2$, $t_{r_1} = 1$, with $\log s_{b_1,b_2 r_1,r_2} = 0.5$, $\sigma_n = 0.2$ and all other hyperparameters initiated as $\theta_i \sim U(0.1,1)$. For comparison, independent Gaussian processes and joint Gaussian Processes were also fitted, with hyperparameters initiated as previously described. Results are summarised in Supplementary Figure 3.



Supplementary Figure 2: Top row: Example posterior GPs fitted to a simple branching process using a three-component branching GP (left), independent GPs (middle) and a joint GP (right). **Middle row:** distribution of mean sum square error (left) and negative log marginal likelihood (right) for BGP, IGP and JGP. **Bottom row:** distribution of mean sum square error (left) and negative log marginal likelihood (right) for BGPs and IGPs only.



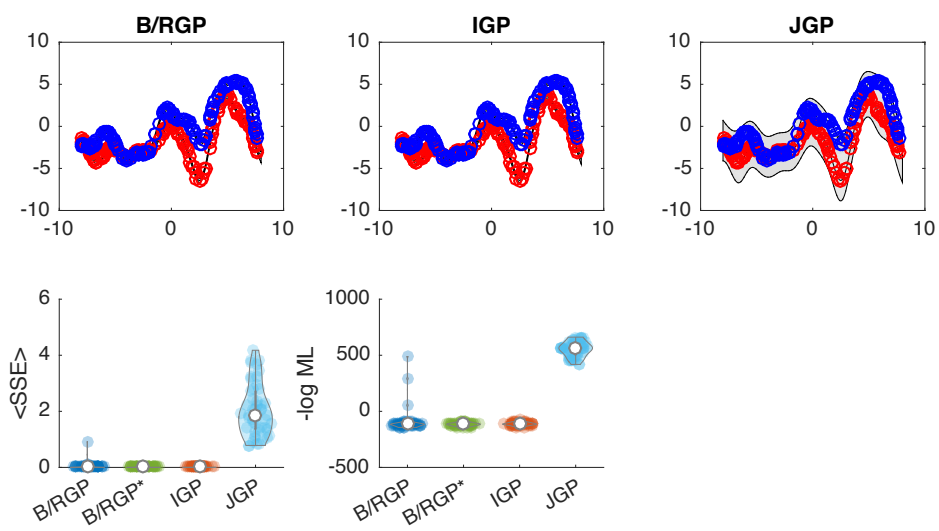
Supplementary Figure 3: Top row: Example posterior GPs fitted to a simple branching/recombination process using a three component B-RGP (left), independent GPs (middle) and a joint GP (right). **Middle row:** distribution of mean sum square error (left) and negative log marginal likelihood (right) for B-RGP, IGPs and JGP. **Bottom row:** distribution of mean sum square error (left) and negative log marginal likelihood (right) for B-RGPs and IGPs only.

To test for robustness to model mismatch, we used B-RGPs on two other datasets. In Experiment 3 a non-branching, three-component process was used to generate data, corresponding to a HGP, $f_0(t) \sim GP\big(0, K_0(t,t')\big)$, $f_1(t) \sim GP\big(f_0(t), K_1(t,t')\big)$, $f_2(t) \sim GP\big(f_0(t), K_2(t,t')\big)$, with squared-exponential covariance functions used throughout. A B-RGP was fitted to the data with hyperparameters initialised as $t_{b_1} = -4$, $t_{r_1} = 4$, $\log s_{b_1,b_2 r_1,r_2} = 0.5$, and $\log \sigma_n = 0.2$, and all other hyperparameters initiated as $\theta_i \sim U(0.1,1)$. In this case there exists a model mismatch between the data, which has no explicit branching or recombination, and the branch-recombinant covariance function used for inference. Nonetheless, we note that informally, if the branch point occurs much earlier than the first data point and the recombinant point occurs much later than the last data point, the behaviour over the range of observations is identical to that of a HGP, $f_1(t) \sim GP\left(0, K_{b_1}(t,t') + K_{b_2}(t,t')\right)$, $f_2(t) \sim GP\left(0, K_{b_1}(t,t') + K_{b_3}(t,t')\right)$. Tuning of the branch/recombination time hyperparameters should therefore allow a good fit over the regions of observation despite the model mismatch. We also considered the explicit limiting case of the B-RGP, with branch/recombination time hyperparameters fixed to arbitrarily low/high values, $t_{b_1} = -40000$ $t_{r_1} = 40000$, with all other hyperparmeters initiated and optimised as before. In Figure 4(top row) we plot example fits to the function using a B-RGP, IGPs and a JGP. In Figure 4 (middle and bottom rows) we indicate the sum of squared errors for the B-RGP and limiting case of the B-RGP (B-RGP*) compared to IGPs and a JGP, as well as the negative log marginal likelihoods. As expected, B-RGPs and IGPs were more accurate than JGPs, due to the increased flexibility to fit the two processes, rather than fitting the general underlying trend. In most cases B-RGPs perform comparably to IGPs, although in a few instances the B-RGP appears to suffer from numerical instability and failed to converge, with the resulting mean SSE and negative log marginal

likelihood distributions heavy tailed and not as favourable as for IGPs. The limiting case of the B-RGP appears to be comparable to the IGPs. These results suggest that B-RGPs offer comparable performance to IGPs, although performance depends on sensible initialisation of hyperparameters.

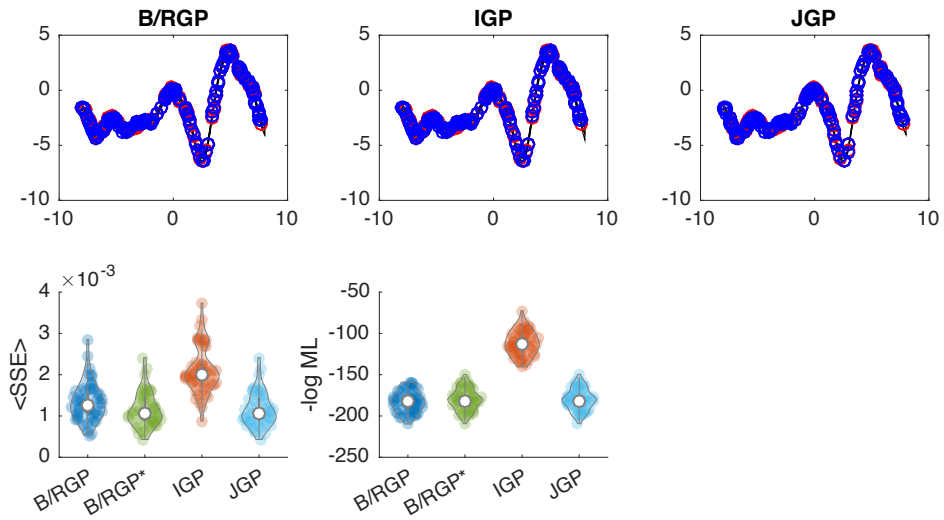
To further evaluate the effect of model mismatch, in Experiment 4 we fit to data from a single, noisy process. Specifically, we used the same three component HGP as in Experiment 3, with noisy observation data generated from the first process only i.e., representing two replicates $y_1 \sim \mathcal{N}(f_1(t), \sigma_n^2 \mathbb{I})$, $y_2 \sim \mathcal{N}(f_1(t), \sigma_n^2 \mathbb{I})$. As before, we fitted the data using a three-component branching-recombinant process, with squared-exponential covariance function assumed for all branches, and hyperparameters initiated $t_{b_1} = -4$, $t_{r_1} = 4$, with all other hyperparameters initiated identically to Experiment 2. Informally, we note that, despite the model mismatch, when branching and recombination both occur much earlier than the first observation, or much later than the last observation, the fit over the range of observations should correspond to that of a JGP with covariance function corresponding to that of the main branch process, $f_{1,2}(t) \sim GP\big(0, K_0(t, t')\big)$. We therefore chose to include the limiting case, where branch and recombination time hyperparameters were fixed arbitrarily early, $t_{b_1} = -50000$ $t_{r_1} = -40000$ with all other hyperparameters initialised and optimised as previously. In Figure 5 (top row) we indicate example fits to the function using a B-RGP, IGPs and a JGP, whilst in Figure 5 (middle and bottom rows) we indicate the distribution for the mean SSE and negative log marginal likelihood. In general both the B-RGP and JGP outperform the IGPs, due to the increased regularisation that can be coded in these models. Both the B-RGP and limiting case of B-RGP offer comparable performance to the JGP.

From preliminary tests on simulated datasets, we note that the B-RGPs represent efficient and flexible models for describing branching and recombination processes. Crucially, all hyperparameters can be optimised directly using type II ML. Therefore branch and recombination time hyperparameters can be tuned, which, due to their limiting behaviour, means that they can gracefully cope with datasets where no branching structure exists, provided hyperparameters are sensibly initialised.



Supplementary Figure 4: Top row: Example posterior GPs fitted to a non-branching hierarchical GP using a three-

component branch-recombinant GP (left), independent GPs (middle) and a joint GP (right). **Bottom row:** distribution of mean sum square error (left) and negative log marginal likelihood (right) for B-RGP, limiting case of a B-RGP (*), IGP and JGP.



Supplementary Figure 5: Top row: Example posterior GPs fitted to a non-branching process (datasets represent replicated samples from a GP) using a three-component branch-recombinant GP (left), independent GPs (middle) and a joint GP (right). **Bottom row:** distribution of mean sum square error (left) and negative log marginal likelihood (right) for B-RGP, limiting case of a B-RGP (*), IGP and JGP.

2.2 TRANSCRIPTIONAL BRANCHING IN ARABIDOPSIS IN RESPONSE TO INFECTION WITH PSEUDOMONAS SYRINGAE

Branching and recombination processes are frequently encountered in datasets involving complex host-pathogen interactions. The initial response to infection is the activation of innate immunity, a highly-conserved response based upon the perception of non-self. Subsequently, pathogens can deliver protein effectors which collectively suppress immunity and collaborate to reconfigure plant metabolism for pathogen nutrition. Thus, initially, the expression of key marker genes will be identically distributed in an infected and uninfected host. Expression patterns will begin to diverge as the host mounts an immune response and as the pathogen activates a unique set of genes for metabolic rewiring of the host. In many cases, the initial immune response is suppressed by the pathogen, potentially driving the expression levels of certain markers back to uninfected levels. Indeed, in *Arabidopsis thaliana* nearly 50% of the transcriptome has been shown to be differentially expressed during a variety of infections (Windram et al. 2012; Lewis et al. 2015). More complex patterns of branching and recombination may exist due to the ongoing evolutionary arms race between pathogens and their hosts (Jones and Dangl 2006).

To evaluate the usefulness of our B-RGPs on a genome scale we therefore classified transcriptional branching patterns in *Arabidopsis thaliana* in response to pathogen infection. As outlined in the main text, we considered six possible branch structures. Briefly, these were: hrpA branches from the control, with DC3000 branching from hrpA at a later point (Group 1), or hrpA and DC3000 independently branch from the control (Group 2), which collectively represent immune response

genes that are targeted by effectors; DC3000, but not hrpA, branches from control (Group 3), representing host susceptibility genes that have been targeted by effectors; hrpA, but not DC3000, branches from the control (Group 4), likely representing immune genes that have been targeted by effectors prior to their natural immune response times; both DC3000 and hrpA jointly branch from the control, but not from one another (Group 5), representing core immune response genes not targeted by effectors; no branching exists (Group 6), representing genes unaffected by plant immunity or pathogen virulence strategies. Example expression patterns of individual genes from each of the six groups are shown in Supplementary Figure 7(a).

For Group 1, we assume that the hrpA-infected time series branches from the mock-infected time series, with the DC3000-infected time series branching from hrpA-infected:

$$
\begin{aligned}
&f_{mock}(t) \sim \mathcal{GP}\big(c, k_{mock}(t,t')\big),\\
&f_{hrpA}(t) \sim \mathcal{GP}\left(f_{mock}(t), CP\left(K_0, k_{hrpA}(t,t')\right)\right),\\
&f_{DC3000}(t) \sim \mathcal{GP}\big(f_{hrpA}(t), CP(K_0, k_{DC300}(t,t'))\big),
\end{aligned}
$$

where observation data were assumed to be a noisy instances of these functions e.g., $y_{mock}(t) = f_{mock}(t) + \varepsilon$. For Group 2, we have hrpA-infected and the DC3000-infected time series independently branching from the mock-infected time series:

$$
\begin{aligned}
&f_{mock}(t) \sim \mathcal{GP}\big(c, k_{mock}(t,t')\big),\\
&f_{hrpA}(t) \sim \mathcal{GP}\left(f_{mock}(t), CP\left(K_0, k_{hrpA}(t,t')\right)\right),\\
&f_{DC3000}(t) \sim \mathcal{GP}(f_{mock}(t), CP(K_0, k_{DC300}(t,t'))).
\end{aligned}
$$

Collectively Groups 1 and 2 should represent immune response genes targeted by effectors, and therefore associated with the onset of disease. For Group 3, we have mock-infected and hrpA-infected datasets drawn from an identical process, with the DC3000-infected branching from this:

$$
\begin{aligned}
&f_{mock,hrpA}(t) \sim \mathcal{GP}\left(c, k_{mock,hrpA}(t,t')\right),\\
&f_{DC3000}(t) \sim \mathcal{GP}\left(f_{mock,hrpA}(t), k_{DC3000}(t,t')\right).
\end{aligned}
$$

This group represents genes not associated with the immune response that are nonetheless targeted by effectors, and may therefore represent those functioning in metabolism. For Group 4, we have mock-infected and DC3000-infected datasets drawn from an identical process, with the hrpA-infected branching from this:

$$
\begin{aligned}
f_{mock,DC3000}(t) &\sim \mathcal{GP}\left(c, k_{mock,DC3000}(t,t')\right),\\
f_{hrpA}(t) &\sim \mathcal{GP}\big(f_{mock,DC3000}(t), CP(K_0, k_{hrpA}(t,t'))\big).
\end{aligned}
$$

These genes likely reflect downstream immune response genes that are targeted very early by effectors, before their usual time of activation. For Group 5, we have hrpA-infected and DC3000-infected datasets drawn from an identical process that branches from mock-infected:

$$f_{mock}(t) \sim \mathcal{GP}\big(c, k_{mock}(t,t')\big),$$

$$f_{hrpA,DC3000}(t) \sim \mathcal{GP}\big(f_{mock}(t), CP(K_0, k_{hrpA,DC3000}(t,t'))\big).$$

These genes represent immune response genes not targeted by effectors. Finally, for Group 6, we have all datasets drawn from an identical process:

$$f_{mock,hrpA,DC3000}(t) \sim \mathcal{GP}\big(c, k_{mock}(t,t')\big),$$

representing genes that are not differentially expressed. Because these datasets correspond to bulk observations from microarrays with well-defined capture times we assumed smooth functions throughout, and therefore, in all cases, the covariance functions were taken to be squared-exponentials e.g., $k_{mock}(t,t') = SE_{\theta_{mock}}(t,t) = \sigma_{mock}^2 \exp\left((t-t')/2{l_{mock}}^2\right)$, where $\theta_{mock} = [{l_{mock}}^2, \sigma_{mock}^2]$ denotes a set of mock dataset-specific hyperparameters. Hyperparameters were optimised to their MAP values. We assumed the following prior distributions: the first branch time was Gamma distributed, $t_{b_1} \sim \Gamma(2,2)$, with the second branch, $t_{b_1} \sim \Gamma(4,2)$, and the change-point transition rate Gaussian distributed, $t_{b_1} \sim \mathcal{N}(0,0.5)$. All other hyperparameters were optimised to their ML values. Finally, we selected the optimal group based on the Bayesian information criterion (BIC).

In Supplementary Figure 6(a) we indicate the branch time between control and hrpA time series using B-RGPs versus that obtained using the Gaussian process two-sample (GP2S) approach of Stegle et al. (2010) (Supplementary Figure 6(b)). Here the GP2S approach appeared to incorrectly identify a peak perturbation time at $t = 0$, before Arabidopsis could mount an immune response. This peak was notably absent in our B-RGP approach. To further gauge the accuracy of our approach, we compared the estimated branch times between hrpA and DC3000 using B-RGPs (Supplementary Figure 6(c)) to that obtained using perturbation times estimated (Yang et al. 2016) (Supplementary Figure 6(d)). The approach of Yang *et al*., (2016) provide 90% confidence intervals for their estimates, and we note that our MAP estimation falls within these bounds in 67% of cases. Of the remaining genes, 27% of our MAP estimates lie to the right of the confidence bounds, and 5% to the left of the confidence bounds, suggesting that our approach has a tendency to estimate later branch times than that of (Yang et al. 2016). This is likely due to differences in the prior distributions over branch times. Indeed, if we plot the estimated branch time using our method versus the PT approach for the 27% of genes noted above, we see a strong correlation (r = 0.8507). Together, these results suggest that, although there is good agreement between the methods for a large fraction of cases, inference of exact branch time in a subset of the observations may be unidentifiable.

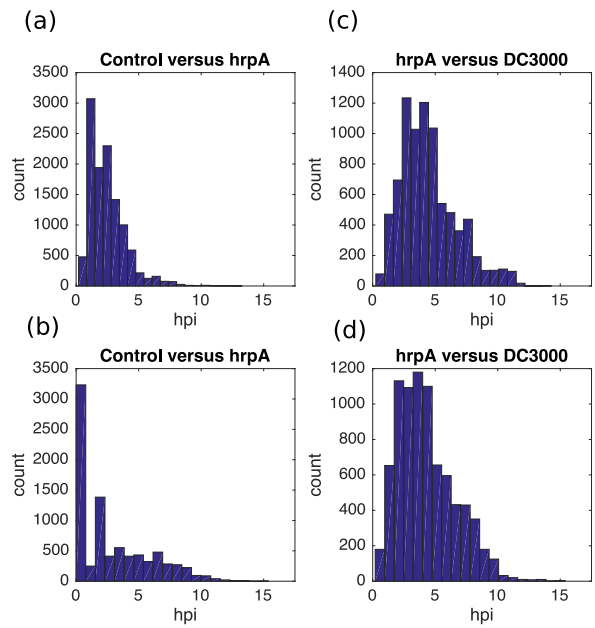
In general, our results indicate that approximately 50% of genes were unperturbed by either the hrpA or DC3000 strains, in agreement with previous studies based on pairwise comparison of the time series using mixtures of Gaussian Processes (Lewis et al. 2015), where 52% of genes were identified as being differentially expressed in control versus hrpA or control versus DC3000. Example expression plots for each of the six groups is shown in Supplementary Figure 7(a).

Since DC3000 is known to subvert the basal immune response of *Arabidopsis*, we hypothesised that the expression of a subset of genes in the DC3000-infected dataset might converge fully back

to control levels following delivery of effectors. To identify such genes, we also fitted a two-component branch-recombinant GP using the control and DC300-infected time series only, using the BIC to distinguish between genes undergoing branching and recombination from those undergoing branching alone. An example branch-recombinant expression profile is shown in Figure 1(b). We note that relatively few genes were identified as having their expression levels fully converge back to control levels, none of which were known targets of effectors or previously implicated in the response to *P. syringae*, suggesting that the full suppression of early immune response genes to control levels is not required for infection to advance.

Gene Ontology analysis identified a number of highly enriched terms across the first five groups (see Supplementary Table 1), suggesting distinct biological functions relating to pathogen response and metabolic reprogramming. Prior to 3 hours, the ontologies represent some of the earliest transcriptional processes targeted by effectors. Consequently, there is a diverse array of GOs represented. Notable are the combination of proteolytic, ribosome, vitamin and amino acid metabolic and transport processes. This is indicative of assembly of the processing machinery to enable effector mediated reprogramming of core cellular processes. Between 3 and 5 hpi the subsequent impact of effectors was evident by the number of GOs identified, with processes associated with nuclear activities, in particular chromatin remodelling, nuclear transport, and transcription being the most highly enriched. Other GO processes such as hormone responses and primary metabolism were, unexpectedly, less abundant. While Lewis et al. (2015) reported evidence for chromatin remodelling, B-RGPs appears to provide much better temporal resolution. As the effector-driven virulence programme proceeds, but prior to bacterial multiplication (5-8 hpi), there is a strong enrichment of terms related to adenyl ribonucleotide binding, reflecting the high energy demands at this transitory phase of the infection process, when *Pseudomonas* effectors have suppressed immune responses and are reconfiguring the metabolism to facilitate pathogen growth

To further investigate the nature of these groups, we looked for enrichment of known targets of effectors of various pathogens (Mukhtar et al. 2011). We first checked for enrichment of non-*Pseudomonas* effectors, hypothesising that the Arabidopsis immune response to *Pseudomonas syringae* should be enriched for targets of other pathogens due to convergent evolution of virulence mechanisms (Mukhtar et al. 2011). Our results are summarised in Figure 10(d, e), which shows that these groups were indeed appropriately enriched. Next, we checked for enrichment of Avr and Hop effectors that are present in various strains of *Pseudomonas syringae*, although not necessarily DC3000 itself, and were, again, appropriately enriched in DC3000-responsive groups (Figure 10(d,f)). We next looked for enrichment of genes with known pathogen-response phenotypes, using TAIR (Huala et al. 2001) to query for genes using the terms 'Pseudomonas', 'Botrytis', and 'Peronospora'. In Supplementary Figure 11 we indicate the frequency of particular pathogen-response genes within various groups, and our results show a distinct enrichment for Pseudomonas-related and Botrytis-related genes amongst the various immune responsive and disease-responsive groups.



Supplementary Figure 6: Histogram of the inferred branch times using B-RGPS. Branch times between mock-infected and hrpA-infected plants are shown for B-RGPs **(a)** and a mixture of Gaussian processes (GP2S) **(b)**. Branch times between hrpA-infected and DC3000-infeced Arabidopsis using B-RGPs **(c)** and Perturbation Time (PT) analysis **(d)**.

2.3 B-RGPS FOR PARTIALLY LABELLED DATA

In our previous examples, inference relied on the existence of explicit branch labels. In some cases, however, branch labels may be incomplete or missing entirely. For example, in a collection of single cell transcriptomics data there may be various cells types, including some that cannot be unambiguously assigned to a particular class *a priori*. Here we can attempt to infer the branch labels, $\mathbf{z}$, using Markov Chain Monte Carlo. Here we assume that a subset of branch labels are know, with the remainder unknown, denoted $\mathbf{z} = [\mathbf{z}^{(\text{labelled})}, \mathbf{z}^{(\text{unlabelled})}]$. Branch labels for the unknown components are first initialised either stochastically or heuristically, and updated via a Gibbs sampler, similar to the usage in Stegle et al. (2010). For an n-component branching process, the unknown label for cell i, is Gibbs sampled conditional on the observation data and branch assignment of all other cells:

$$P(z_i = I | \boldsymbol{z} \backslash z_i, \boldsymbol{y}, \boldsymbol{t}, \theta) = \frac{P(\boldsymbol{y} | z_i = I, \boldsymbol{z} \backslash z_i, \boldsymbol{t}, \theta)}{\sum_k P(\boldsymbol{y} | z_i = k, \boldsymbol{z} \backslash z_i, \boldsymbol{t}, \theta)},$$

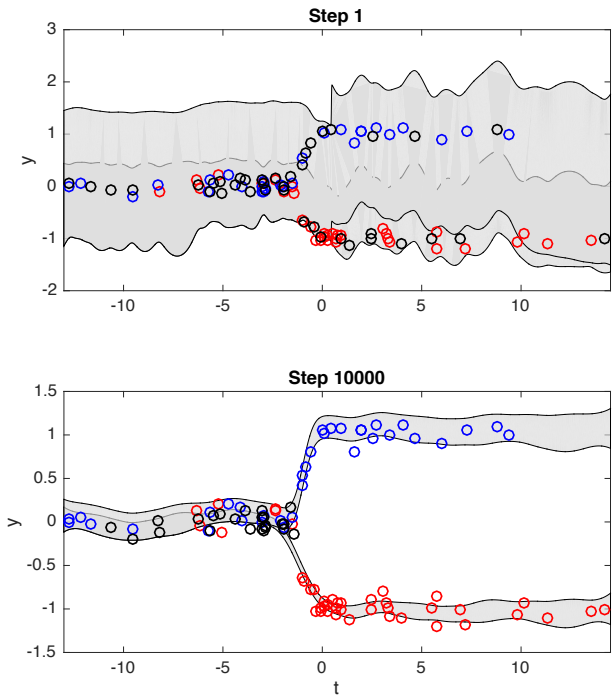
with the subsequent update of hyperparameters using Hybrid Monte Carlo conditional on all branch labels:

$$\theta \sim P(\theta | \boldsymbol{t}, \boldsymbol{z}, \boldsymbol{y}).$$

To test the accuracy of our B-RGPs on unlabelled data we generated observations from the simple branching process used in Experiment 1 of Supplementary Section 2. Here we first generated a set of test input locations, $t \sim \mathcal{N}(0, 5\mathbb{I})$, with observation data generated as noisy instances of the

process. We then attempted to infer branch labels and hyperparameters within an MCMC scheme, with labels updated via Gibbs sampling, and hyperparameters sampled using Hybrid Monte Carlo, as outlined above. A subset of data points, n, were (initially) randomly assigned to either branch with equal probability, where $n/N \in [0.25, 0.5, 0.75, 1]$ and N indicates the total number of observations, with the remaining data points assigned labels heuristically, according to whether the value of that particular data point was above or below the overall mean. Five randomly initiated runs were used, with 10,000 steps in the MCMC chain, and the first 1000 discarded for burn-in. The fraction of correctly classified data points is summarised in Supplementary Table 2, and an example fit shown in Supplementary Figure 8. We note that, in this example, although the initial heuristic assignment already classified the data points reasonably well, the accuracy tended to increase following MCMC updates, suggesting B-RGPS can be used for partially labelled datasets.

An alternative approach for initialising branch labels was also trialled. Here we generated 1000 sample process from a two-component branching process with randomly initialised hyperparameters. Each data point was then hard assigned to the closest branch, and the optimal initial configuration selected based on the smallest SSE over those 1000 samples. Results are again summarised in Supplementary Table 2, with an example fit is shown in Supplementary Figure 8.



Supplementary Figure 8: Inferring the B-RGPs for unlabelled data. Initially data points are assigned branch labels heuristically, with labels updated using Gibbs samples. After 10,000 iterations the B-RGP recovers the branching structure inherent to the data.

3 PSEUDO-TEMPORAL ORDERING OVER BRANCHING PROCESSES

Single cell transcriptomics datasets typically have low temporal resolution with many individual cells measured over relatively few developmental stages. However, due to the heterogeneity of individual cells, these low resolution time series can often reflect a continuum of states, essentially representing a much high temporal resolution provided that the cells can be suitably ordered. A

key goal in single cell transcriptomic analysis is to order cells along a developmental axis that reflects the relative progression of each cell (pseudo-time ordering). For a cells progressing through a single biological process, with capture times $t_c = (t_1, \dots, t_M)$, the aim is to infer the corresponding set of pseudotimes, $t_p = (\tau_1, \dots, \tau_M)$. In general, this can be achieved using a Gaussian process latent variable models (Lawrence 2003, 2005), and recent studies by Reid and Wernisch (2016) have demonstrated how to use GPLVMs with informative (Gaussian) priors over the pseudo-times. Within these models, each gene is assigned an independent Gaussian process prior over the pseudo-times, with the composite likelihood written as:

$$\mathcal{L}(\boldsymbol{t}_p) = P(\boldsymbol{t}_p|\boldsymbol{t}_c) \prod_{i=1}^{n_g} \int P(\boldsymbol{y}_i|\boldsymbol{f}, \boldsymbol{t}_p, \boldsymbol{z}) P(\boldsymbol{f}|\boldsymbol{t}_p, \boldsymbol{z}) d\boldsymbol{f},$$

where $\boldsymbol{y}_i$ represents the vector of expression for gene i over all cells, and $P(\boldsymbol{t}_p|\boldsymbol{t}_c)$ denotes the prior probability of the cells having pseudo-times, $\boldsymbol{t}_p$ conditional on their capture times, $\boldsymbol{t}_c$, which is typically assumed to be Gaussian distributed: $\boldsymbol{t}_p|\boldsymbol{t}_c \sim \mathcal{N}(\boldsymbol{t}_c, \sigma^2\mathbb{I})$. When there exists a branching process with capture times, $t_c = (t_1, \dots, t_M)$, and branch-labels, $z_c = (z_1, \dots, z_M)$, each gene can similarly be assigned an independent B-RGP prior over the pseudo-times, and the composite likelihood is thus identical to that outlined above.

3.1 PSEUDO-TEMPORAL ORDERING OF ARABIDOPSIS DATA FOLLOWING INNOCULATION WITH BOTRYTIS CINEREA

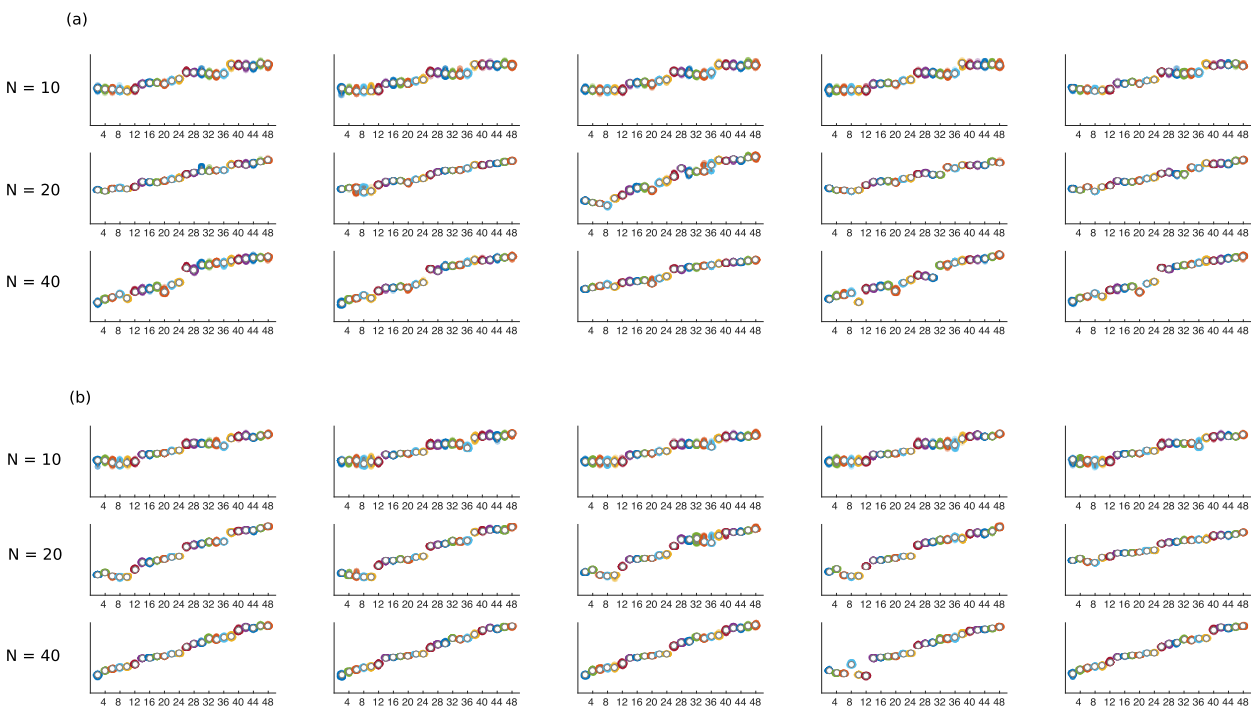
To assess the ability of our approach to pseudo-temporally order data over a branching process, we benchmarked using microarray time series data measuring the *Arabidopsis thaliana* transcriptional response to the necrotrophic bacteria, *Botrytis cinerea* (Windram et al. (2012); GEO GSE39597). This dataset consists of two time-series: (i) a control time series, detailing changes in gene expression in Arabidopsis over a 48-hour period at two-hourly intervals; and (ii) a time-matched infection series, in which Arabidopsis has been inoculated with *B. cinerea*. The infection time series has previously been used to benchmark GPLVMs for pseudo-time ordering (Reid and Wernisch 2016). Here we performed pseudo-time ordering over a branching process using both the control and infection time series simultaneously. As in Reid and Wernisch (2016), we first grouped the individual measurements into four groups containing six consecutive time points, artificially reducing the temporal resolution of each time series dataset from 24 time points to 4. We then attempted to recapture the correct ordering of cells using an increasing number of randomly selected genes within the GPLVM. Here we selected 10, 20 and 40 genes respectively from the set of 150 previously used by Reid and Wernisch (2016). In doing so, we aimed to identify how the accuracy of the model changed with increased number of gene observations. We used 5 different randomly initialised runs, with 10,000 samples in the MCMC chain, discarding the first 5,000 steps for burn-in. For each independently initialised run, a separate random selection of genes was chosen, in order to gauge variability in the inferred ordering.

Since these time series were generated using microarrays, with measurements representing fluorescence based on populations of cells, we assumed a smoothly varying process, and thus used a two-component branching process consisting of squared-exponential covariance functions. Hyperparameters were initially fitted using maximum likelihood estimates based on the low-

resolution capture times, $\theta \leftarrow \arg_\theta \max \mathcal{L}\left(\theta | \boldsymbol{t}_p = \boldsymbol{t}_c, \boldsymbol{y}\right)$. Length-scale, process-variance and noise hyperparameters were then fixed throughout the inference. For the change-point time hyperparameters, we assumed a univariate smoothed box prior distribution with linear decay in the log domain, $P(\boldsymbol{t}|a,b) = \sigma(\eta(t-a))(1-\sigma\left(\eta(t-b)\right))$, where $\sigma(z) = 1/(1+\exp(-z))$, and $a = 0$, $b = 33$, and $\eta = 33$. Within the MCMC algorithm, we considered the following updates:

1. A Gaussian perturbation of a subset of pseudo-times.
2. A random swap of pseudo-times between adjacent data-points.
3. A global swap of pseudo-times between two randomly selected data points.
4. Every 10 steps we sampled the change-point hyperparameters via HMC using 100 steps.

In Supplementary Figure 9 we plot the inferred pseudo-time ordering for control and infection datasets versus the true ordering of data as the number of genes used within the GPLVM increased, with $N \in (10,20,40)$. We note that, in general, that the accuracy of the models appears to increase as the number of genes increased. For the case $N = 40$, we noted that 4 out of 5 replicates were perfect ordered in the infected time series, with good ordering observed the control time series. Consequently, when pseudo-temporally ordering single cell data, we used approximately 40 genes.



Supplementary Figure 9: Pseudo-temporal ordering (y-axis) versus true-ordering (x-axis) of data using based upon *Arabidopsis thaliana* response to *Botrytis cinerea*. Here we indicate the ordering for the mock time series **(a)** and the infected time series **(b)**.

4 COMPLEX BRANCHING DURING EARLY EMBRYO DEVELOPMENT IN HUMANS

Single cells from pre-implantation embryos, PGCs, somatic cells and ESCs were initially pseudo-temporally ordered over a two-component branching process as outlined in the main text. Once the pseudo-temporal ordering of cells had been established we looked for more complex branch-recombinant structures. Here we considered the dynamics of PGCs, somatic cells, and ESCs. Each of the three groups was randomly assigned pre-implantation cells with equal probability, except for the ESCs, which partially overlapped with blastocyst stage cells; here we instead randomly assigned pre-blastocyst cells. To capture the dynamics necessary to drive ESCs towards week 4 PGC fate, we also assigned half of the male hPGCs to the ESCs class. For model 1, we assumed:

$$f_{PGC}(t) \sim \mathcal{GP}\big(c, k_{PGC}(t,t')\big),$$
$$f_{soma}(t) \sim \mathcal{GP}\left(f_{PGC}(t), CP\big(K_0, k_{PGC}(t,t')\big)\right),$$

$$f_{ESC}(t) \sim \mathcal{GP}\big(f_{PGC}(t), CP(CP\big(K_0, k_{DC300}(t,t')\big)K_0)\big),$$

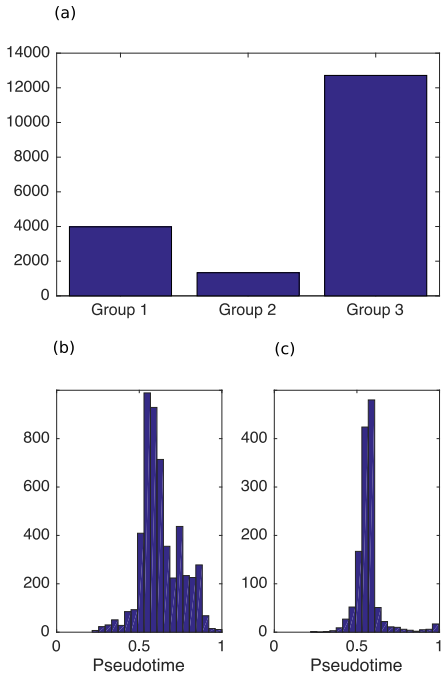
representing genes in which somatic cells diverged from PGCs, and where ESCs diverged from the *in vivo* dynamics of PGC specification before recombining at around week 4. For model 2, we assumed:

$$f_{PGC}(t), f_{ESC}(t) \sim \mathcal{GP}\big(c, k_{PGC}(t,t')\big),$$
$$f_{soma}(t) \sim \mathcal{GP}\left(f_{PGC}(t), CP\big(K_0, k_{PGC}(t,t')\big)\right).$$

This model represented genes that showed divergence between PGCs and soma, with ESCs identically distributed to PGCs. Finally, for model 3 we assumed:

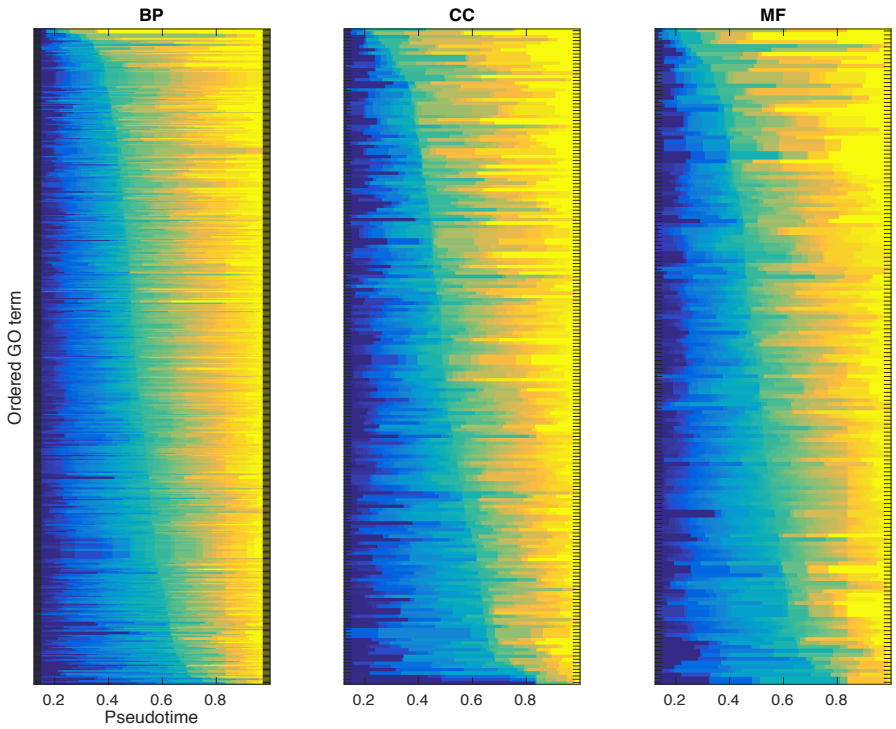
$$f_{PGC}(t), f_{soma}(t), f_{ESC}(t) \sim \mathcal{GP}\big(c, k_{PGC}(t,t')\big),$$

representing genes that showed no divergence between PGCs, soma or ESCs. Because these datasets represent single-cell measurements that might be intrinsically less smooth than the microarray datasets, we assumed a Matérn covariance function throughout. As in previous analyses, we optimised hyperparameters to their MAP values and used the BIC to select the branching structure for each gene. The frequency of the three groups is indicated in Supplementary Figure 10(a), which shows the majority of genes were not differentially expressed i.e., non-branching. A histogram of the timing of branching is indicated in Supplementary Figure 10(b, c), and shows that the majority of genes that diverge between PGCs and soma cells do so between blastocyst stage and week 4, as expected.

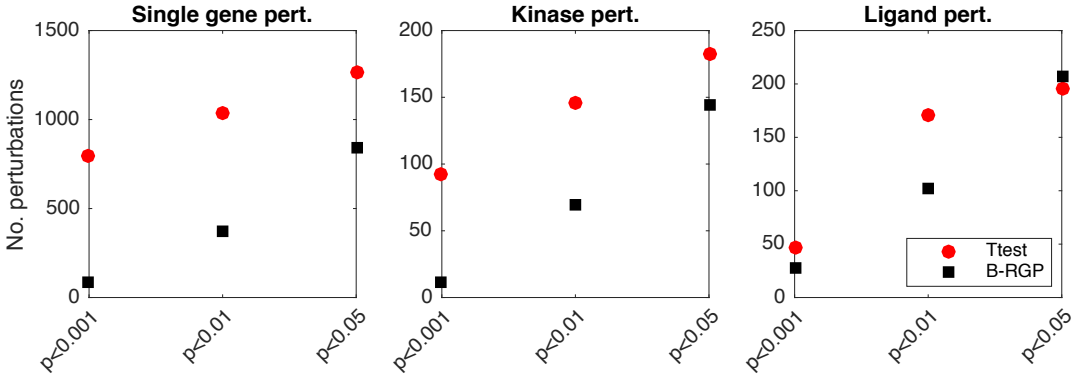


Supplementary Figure 10: (a) Frequency of the different groups. **(b)** Timing of branching between PGC and somatic cells. **(c)** Time of recombination between ESCs and inferred *in vitro* dynamics.

Following GO analysis on the individual groups using permissive p-values ($p<0.1$ following multiple hypothesis test correction), we ordered terms by the pseudo-time at which 50% of their associated genes had branched, as previously done in Yang et al. (2016). In Figure 16 we indicate a heatmap representation of this ordering, where the x-axis represents the pseudo-time, the y-axis represents individual GO terms, and the colour indicates the fraction of genes that had branched. From these heatmaps we can see a continuum of responses. An associated table containing the full list of terms associated with the y-axis is available in Supplementary Table 4.



Supplementary Figure 11: Individual GO terms for genes that were up-regulated in PGCs versus soma, according to the timing at which at least 50% of the genes associated with each term had branched. See also associated Supplementary Table 4.



Supplementary Figure 12: The number of enriched perturbations from Enrichr using different p-value thresholds (single gene perturbations, kinase perturbations, and ligand perturbations). Here we indicate the number of perturbations identified via differential expression analysis of PGC expression versus soma expression using a Student's t-test, or following selection of early branched genes using B-RGPs. The selection of early events identifies far fewer putative perturbations than simple differential expression analysis, allowing for more focused experiments.

4.1. ANALYSIS OF *SOX17* OVEREXPRESSION LINES

We identified differentially expressed genes in hESC lines constitutively over-expressing *SOX17* (Seguin et al. (2008); GSE10809). Differential expression was evaluated using a Student's t-test

on log2 fluorescence versus parental lines using a cut-off of $p<0.03$ and log2 fold change >2 (similar to Seguin et al. (2008)).

4.2. GENE SET ENRICHMENT ANALYSIS

Gene set based enrichment analysis was used to identify significantly enriched terms using Enrichr database. Here we focused on enrichment of GO 2017 and KEGG 2016 terms using permissive p-values ($p<0.1$ following multiple hypothesis testing correction). Using Enrichr we also looked for enrichment of Kinase Perturbations from GEO, Single Gene Perturbations from GEO, and Ligand Perturbations from GEO, using p-value cut-off of $p<0.01$ (following multiple hypothesis testing correction).

Enrichment of plant ontologies was evaluated using Plant Gene Set Enrichment Analysis (Yi, Du, and Su 2013), for the following ontologies: Biological Process (BP), Cellular Components (CC), Molecular Function (MF), KEGG, and PlantCyc. In supplementary Table 1 we report all terms with FDR <0.05.

5 EXPERIMENTAL PROCEDURES

5.1 CELL CULTURE AND PGCLC DIFFERENTIATION

hESCs (WIS2-NANOS3-T2A-tdTomato (N3-tdTom, Kobayashi et al. (2017)) were cultured as in (Irie et al. 2015)on irradiated mouse embryonic fibroblasts (MEFs) (GlobalStem) in 4i medium (Supplementary Table 7) or modifications therefrom (Supplementary Table 8). Media were replaced every day. hESCs were passaged by single-cell dissociation using 0.25% Trypsin-EDTA (GIBCO). 10 μM ROCK inhibitor (Y-27632, TOCRIS) was added for 24 hours after passaging.

To induce PGCLCs (Irie et al. 2015), hESCs were trypsinized, filtered and plated into ultra-low cell attachment U-bottom 96-well plates (Corning, 7007) at 4,000 cells/well density in 100 μl PGCLC medium (Supplementary Table 9) The plate was centrifuged at 300g for 3 minutes and placed into a 37 °C 5% CO_2 incubator until embryoid body (EB) collection for downstream analysis. Reporter fluorescence intensities were monitored daily throughout differentiation using Olympus IX71 microscope.

Supplementary Table 7. Competent (4i) medium composition.

Component	Final Concentration	Supplier
Knockout DMEM	-	GIBCO
Knockout serum replacement (KSR)	20%	GIBCO
L-glutamine	2 mM	GIBCO
Nonessential amino acids	0.1 mM	GIBCO
2-mercaptoethanol	0.1 mM	GIBCO
Penicillin-Streptomycin	100 U/ml (Penicillin)	GIBCO

	0.1 mg/ml (Streptomycin)	
Human LIF	20 ng/ml	Stem Cell Institute (SCI)
bFGF	8 ng/ml	SCI
TGF-β1	1 ng/ml	Peprotech
CHIR99021 (CH)	3 μM	Miltenyi Biotec
PD0325901 (PD)	1 μM	Miltenyi Biotec
SB203580 (SB)	5 μM	TOCRIS bioscience
SP600125 (SP)	5 μM	TOCRIS bioscience

Supplementary Table 8. Modifications of the 4i medium used in the competence screen.

Medium	Recipe
4i-PD03	4i as in Supp. Table 7 without PD0325901
4i+PD17	4i as in Supp. Table 7 supplemented with indicated concentrations of PD173074 (TOCRIS bioscience)
4i-PD03+PD17	4i-PD as above supplemented with indicated concentrations of PD173074 (TOCRIS bioscience)
4i+LY	4i as in Supp. Table 7 supplemented with 10 μM LY294002 (Sigma)
4i-PD+LY	4i-PD as above supplemented with 10 μM LY294002 (Sigma)
4i+IWR	4i as in Supp. Table 7 supplemented with 2.5 μM IWR-1 (Sigma)
4i-CH	4i as in Supp. Table 7 without CHIR99021
4i-CH+IWR	4i as in Supp. Table 7 without CHIR99021 supplemented with 2.5 μM IWR-1 (Sigma)
4i-SB	4i as in Supp. Table 7 without SB203580
4i-SP	4i as in Supp. Table 7 without SP600125
4i-LIF	4i as in Supp. Table 7 without human LIF
4i+JAKi	4i as in Supp. Table 7 supplemented with 1μM CAS457081-03-7

Supplementary Table 9. PGCLC induction medium composition

Component	Final Concentration	Supplier
Glasgow's MEM (GMEM)	-	GIBCO
KSR	15%	GIBCO
L-glutamine	2 mM	GIBCO
Nonessential amino acids	0.1 mM	GIBCO
2-mercaptoethanol	0.1 mM	GIBCO

Penicillin-Streptomycin	100 U/ml (Penicillin) 0.1 mg/ml (Streptomycin)	GIBCO
Sodium pyruvate	1 mM	Sigma
BMP2	500 ng/ml	SCI
Human LIF	1 μg/ml	SCI
SCF	100 ng/ml	R&D Systems
EGF	50 ng/ml	R&D Systems
ROCK inhibitor (Y-27632)	10 μM	TOCRIS bioscience

5.2. Flow Cytometry

D4 or D5 EBs were washed in PBS and dissociated with 0.25% Trypsin-EDTA for 15 min at 37°C. Cells were resuspended in FACS buffer (3% FBS in PBS) and incubated with antibodies specified in Supplementary Table 10. After washing with FACS buffer, the cells were recorded on BD LSR Fortessa. Data were analysed using FlowJo (Tree Star).

Supplementary Table 10. Antibodies used for flow cytometry

Name	**Dilution**	**Manufacturer**
Alkaline Phosphatase PerCP-Cy5.5 #561508	2.5 μl/assay	BD Pharmingen
Alkaline Phosphatase AF647 #561500	2.5 μl/assay	BD Pharmingen
CD38 PerCP-Cy5.5 #303522	2.5 μl/assay	BioLegend
CD38 AF647 #303514	2.5 μl/assay	BioLegend

5.3. Real-Time Quantitative RT-PCR

Total RNA was extracted from unsorted hESCs using RNeasy Mini Kit (QIAGEN). cDNA was synthesized using QuantiTect Reverse Transcription Kit (QIAGEN). qPCR was performed on a QuantStudio 12K Flex Real-Time PCR machine (Applied Biosystems) using SYBR Green JumpStart Taq ReadyMix (Sigma) and human-specific primers (Supplementary Table 11). The $\Delta\Delta Ct$ method was used for quantification of gene expression. Three technical replicates were used for each biological replicate.

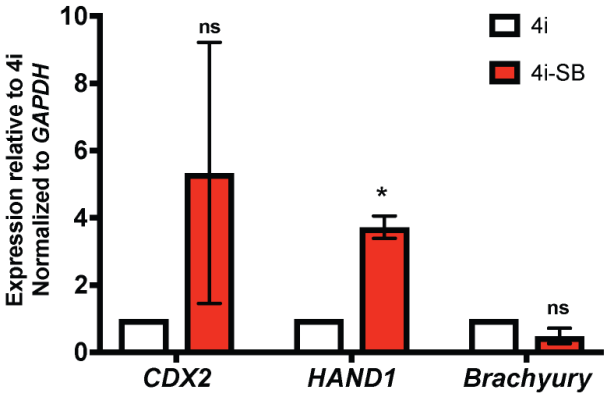
5.4. Dominant-negative ERK expression

For inducible expression of dominant-negative ERK1 (DN-ERK), a cDNA from pFLAG-CMV-hErk1(K71R) (Addgene plasmid #49329) encoding human ERK1 with K71R mutation (Robbins et al. 1993) was cloned (into a doxycycline (dox)-inducible PB-TRE-3G vector (Kobayashi et al.

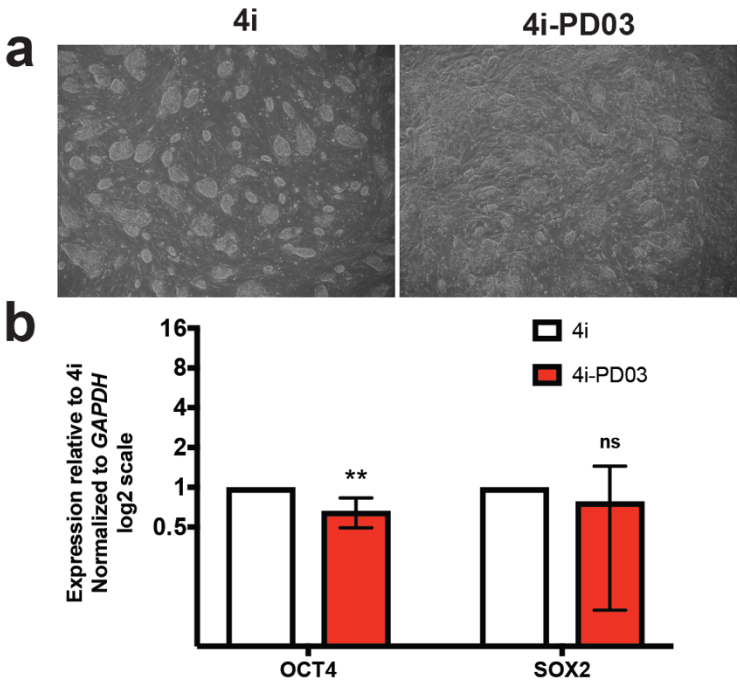
2017) to yield pPB-TRE-3G-DN-ERK-IRES-EGFP. The plasmid was generated using In-Fusion cloning (Clontech) according to manufacturer's recommendations. Primers used for cloning are specified in Table 8. N3tdTom hESCs were co-lipofected with 2 µg PB-TRE-3G-DN-ERK-IRES-EGFP, 2.5 µg PBase and 0.5 µg pPBTET3G-Neo (Kobayashi et al. 2017). Lipofection was performed using OptiMEM (GIBCO) and Lipofectamin 2000 (Invitrogen) according to manufacturer's recommendations.

Supplementary Table 11. Oligos used in the study.

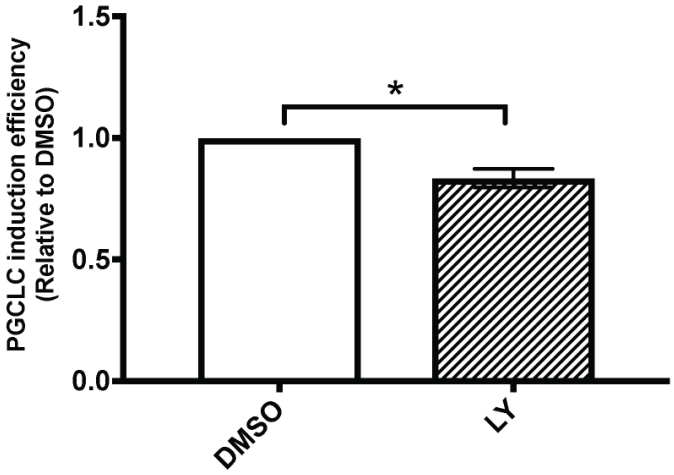
Name	Used for	5'-3' sequence
SOX2_h_F	qPCR	GGGAAATGGGAGGGGTGCAAAAGAGG
SOX2_h_R	qPCR	TTGCGTGAGTGTGGATGGGATTGGTG
OCT4_h-m_F	qPCR	GCTGGAGCAAAACCCGGAGG
OCT4_h-m_R	qPCR	TCGGCCTGTGTATATCCCAGGGTG
CDX2_h_F	qPCR	TCACCATCCGGAGGAAAGCC
CDX2_h_R	qPCR	CTCTCCTTTGCTCTGCGGTT
HAND1_h_F	qPCR	CGTGAGAGCAAGCGGAAAAG
HAND1_h_R	qPCR	AATCCTCTTCTCGACTGGGC
In-Fusion pPBTRE3G-hERK1_F	Cloning	ccctcgtaaagtcgaCGCCGCCACCatgGCGGCGGCGGCGGCTCAG
In-Fusion IRES2-hERK1_R	Cloning	GGGGAGGGAGAGGGGCTAGGGGGCCTCCAGCACTCCGGGCTG
In-Fusion pPBTRE3G-EGFP_R	Cloning	gggccccggtgtcgaCTTACTTGTACAGCTCGTCCATGCCG



Supplementary Figure 13. Withdrawal of p38 inhibitor from 4i results in hESC differentiation. hESCs were cultured in respective media for 4 passages and collected for qPCR analysis. Data are shown as mean (of three technical replicates each) ± SD of 2 independent experiments. * $p \leq 0.05$, ns: not significant ($p > 0.05$), Holm-Sidak t-test.



Supplementary Figure 14. Withdrawal of MEK inhibitor from 4i results changes hESC colony morphology and induces expression of endoderm-related markers. **(a)** Representative pictures of hESCs cultured in complete 4i or "4i-PD03" media. **(b)** qPCR showing pluripotency and endoderm marker genes expression in bulk hESCs cultured in 4i and "4i-PD03" media. Data are shown as mean (of three technical replicates each) ± SD of 5 independent experiments. * $p \leq 0.05$, ** $p \leq 0.01$, ns: not significant ($p > 0.05$), Holm-Sidak t-test.



Supplementary Figure 15. Inhibition of PI3K during differentiation does not abrogate PGCLC specification. Data are shown as mean ± SD of 2 independent experiments. * $p \leq 0.05$, Holm-Sidak t-test.

REFERENCES

Hensman, J., N.D. Lawrence, and M. Rattray. 2013. 'Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters', *BMC Bioinformatics*, 14: 252.

Huala, E., AW. Dickerman, M. Garcia-Hernandez, D. Weems, L. Reiser, F. LaFond, D. Hanley, D. Kiphart, M. Zhuang, W. Huang, LA. Mueller, D. Bhattacharyya, D. Bhaya, B.W. Sobral, W. Beavis, D.W. Meinke, C.D. Town, C. Somerville, and S. Y. Rhee. 2001. 'The arabidopsis information resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant', *Nucleic Acids Research*, 29.

Irie, N., L. Weinberger, W. W. Tang, T. Kobayashi, S. Viukov, Y. S. Manor, S. Dietmann, J. H. Hanna, and M. A. Surani. 2015. 'SOX17 is a critical specifier of human primordial germ cell fate', *Cell*, 160: 253-68.

Jones, J. D., and J. L. Dangl. 2006. 'The plant immune system', *Nature*, 444: 323-9.

Kobayashi, T., H. Zhang, W.W.C. Tang, N. Irie, S. Withey, Klisch. Doris., A. Sybirna, S. Dietmann, D.A. Contreras, R. Webb, C. Allegrucci, R. Alberio, and M.A. Surani. 2017. 'Principles of early human development and germ cell program from conserved model systems', *Nature*.

Lawrence, N.D. 2003. "Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data." In *Advances in Neural Information Processing Systems*, 5.

———. 2005. 'Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models', *Journal of Machine Learning Research*, 6: 1783–816.

Lewis, L. A., K. Polanski, M. de Torres-Zabala, S. Jayaraman, L. Bowden, J. Moore, C. A. Penfold, D. J. Jenkins, C. Hill, L. Baxter, S. Kulasekaran, W. Truman, G. Littlejohn, J. Prusinska, A. Mead, J. Steinbrenner, R. Hickman, D. Rand, D. L. Wild, S. Ott, V. Buchanan-Wollaston, N. Smirnoff, J. Beynon, K. Denby, and M. Grant. 2015. 'Transcriptional Dynamics Driving MAMP-Triggered Immunity and Pathogen Effector-Mediated Immunosuppression in Arabidopsis Leaves Following Infection with Pseudomonas syringae pv tomato DC3000', *Plant Cell*, 27: 3038-64.

Lloyd, J.R, D. Duvenaud, R. Grosse, J.B. Tenenbaum, and Z. Ghahramani. 2014. 'Automatic Construction and Natural-Language Description of Nonparametric Regression Models', *arXiv:1402.4304*.

Mukhtar, M.S., A. Carvunis, M. Dreze, P. Epple, J. Steinbrenner, J. Moore, M. Tasan, M. Galli, T. Hao, M.T. Nishimura, S.J. Pevzner, S.E. Donovan, L. Ghamsari, B. Santhanam, V. Romero, M.M. Poulin, F. Gebreab, B.J. Gutierrez, S. Tam, D. Monachello, M. Boxem, C.J. Harbort, N. McDonald, L. Gai, H. Chen, Y. He, European Union Effectoromics Consortium, J. Vandenhaute, F.P. Roth, D.E. Hill, J.R. Ecker, M. Vidal, J. Beynon, P. Braun, and J.L. Dangl. 2011. 'Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network', *Science*, 333: 596–601.

Rasmussen, C. E., and C. K. Williams. 2006. *Gaussian processes for machine learning* (MIT Press).

Reid, J.E., and L. Wernisch. 2016. 'Pseudotime estimation: deconfounding single cell time series', *Bioinformatics*, 32: 2973-80.

Robbins, D. J., E. Zhen, H. Owaki, C. A. Vanderbilt, D. Ebert, T. D. Geppert, and M. H. Cobb. 1993. 'Regulation and properties of extracellular signal-regulated protein kinases 1 and 2 in vitro', *J Biol Chem*, 268: 5097-106.

Seguin, C. A., J. S. Draper, A. Nagy, and J. Rossant. 2008. 'Establishment of endoderm progenitors by SOX transcription factor expression in human embryonic stem cells', *Cell Stem Cell*, 3: 182-95.

Stegle, O., K.J. Denby, E.J. Cooke, D.L. Wild, Z. Ghahramani, and K.M. Borgwardt. 2010. 'A Robust Bayesian Two-Sample Test for Detecting Intervals of Differential Gene Expression in Microarray Time Series', *JOURNAL OF COMPUTATIONAL BIOLOGY*, 17: 355–67.

Windram, O., P. Madhou, S. McHattie, C. Hill, R. Hickman, E. Cooke, D. J. Jenkins, C. A. Penfold, L. Baxter, E. Breeze, S. J. Kiddle, J. Rhodes, S. Atwell, D. J. Kliebenstein, Y. S. Kim, O. Stegle, K. Borgwardt, C. Zhang, A. Tabrett, R. Legaie, J. Moore, B. Finkenstadt, D. L. Wild, A. Mead, D. Rand, J. Beynon, S. Ott, V. Buchanan-Wollaston, and K. J. Denby. 2012. 'Arabidopsis defense against Botrytis cinerea: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis', *Plant Cell*, 24: 3530-57.

Yang, J., C.A. Penfold, M.R. Grant, and M. Rattray. 2016. 'Inferring the perturbation time from biological time course data', *Bioinformatics*, 32: btw329.

Yi, X., Z. Du, and Z. Su. 2013. 'PlantGSEA: a gene set enrichment analysis toolkit for plant community', *Nucleic Acids Res*, 41: W98-103.