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

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

[1]Wellcome/CRUK Gurdon Institute, University of Cambridge, [2]Wellcome/MRC Cambridge Stem Cell Institute, University of Cambridge, [3]Physylogy, 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.

*To whom correspondence should be addressed. [§]These authors contributed equally.

Contact: c.penfold@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 are reversible, particularly during the onset of disease. Single cell transcriptomics provides a rich resource to explore the temporal progression of bifurcations in gene activity and are therefore useful for elucidating the mechanisms of cell fate decisions, provided that the cells can be suitably ordered over a developmental axis. Most methods for inferring this ordering have been developed for heterogeneous populations of cells collected at single time points, with few approaches specifically designed with structured data in mind, such as single cell time-series data. Recent advances based on Gaussian process latent variable models (GPLVMs) address this by allowing the incorporation of prior information, such as capture time, to yield more accurate ordering of cells. However, such approaches do not allow the ordering of cells over developmental processes with more than one branch. Here we develop a pseudotime approach that allows the ordering of cells over developmental trajectories with arbitrary numbers of branches, in the form of branch-recombinant Gaussian process latent variable models (B-RGPLVM). We use first demonstrate the advantage of our approach compared to existing pseudotime algorithms. Subsequently, we use B-RGPLVMs to infer cell ordering that occurs during early human development as primordial germ cells (PGCs), the precursors of sperm and egg, and somatic cells diverge 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, as well as transcription factors and epigenetic modifiers. By concentrating on the earliest branched signalling events, we identified an antagonistic role for FGF receptor (FGFR) signalling pathway in the acquisition of competence for human PGC fate. We experimentally validate our predictions using pharmacological blocking of FGFR or its downstream effectors (MEK, PI3K and JAK), and demonstrate enhanced competency for PGC fate *in vitro*. Thus, B-RGPLVMs 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. **Software is available from:** https://github.com/cap76/BranchingGPs

1 Introduction

During embryogenesis, individual cells undertake a series of cell fate decision to form a complete embryo comprised of myriad cell types. Each cell fate decision can be thought of in terms of a bifurcation (Poincaré 1885), with the expression levels of key marker genes diverging between the two cell fates, epitomising a branching process. Reciprocal behaviour is encountered in recombination processes, where two or more statistical processes converge, such as when two or more cell types share a common terminal fate. A key challenge is to infer the mechanisms and inductive signals of these decision-making processes by identifying the ordering of branching events. A striking example of transcriptional branching occurs in early human development, when the inner cell mass (ICM) of the blastocyst segregates into hypoblast and epiblast, with the latter subsequently differentiating into ectoderm, mesoderm, and endoderm during the process of gastrulation (Irie, Tang, and Azim Surani 2014). At around this time, circa weeks 2-3 in humans, primordial germ cells (PGC), the embryonic precursors of gametes, are also specified (Irie et al. 2015; Kobayashi et al. 2017). Later, at weeks 5-6, specified PGCs undergo comprehensive epigenetic reprogramming that includes 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 precursors to the germline, PGCs are ultimately responsible for passing on genetic and epigenetic information to the next generation, so understand-

ing the mechanisms of their specification and development is of significant practical use.

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 fate in humans (Irie et al. 2015; Tang et al. 2015), making their specification distinct from that in mice, which involves the combined 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 development (Sasaki et al. 2015a; Irie et al. 2015). Interestingly, human ESCs in conventional cultures have low competence to form PGCLCs, but hESCs grown in a specially formulated "competent" medium gain competence for PGC fate and respond to BMP signalling giving rise to PGCLCs that presumably resemble premigratory *in vivo* PGCs (Irie et al. 2015).

Despite recent advances in understanding of PGC development (Irie et al. 2015; Tang et al. 2015) , the mechanisms allowing human PGC competence acquisition remain poorly understood. Single cell RNA-seq (scRNA-seq) datasets provide unique opportunities to dissect the dynamics of cell fate decisions. In the past scRNA-seq datasets tended to be of low temporal resolution, albeit with observations in many cells, which reflect a continuum of developmental states provided the cells can be appropriately ordered (Yan et al. 2013; Guo et al. 2015; Petropoulos et al. 2016; Borensztein et al. 2017; Huang et al. 2017). Most approaches for pseudotemporal ordering of scRNA-seq datasets rely on manifold learning: that is, a preliminary dimensionality reduction, with cells ordered over the reduced dimensional space, typically by utilising curve fitting or graph-theoretic approaches (Trapnell et al. 2014; Bendall et al. 2014; Marco et al. 2014; Ji and Ji 2016; Setty et al. 2016). Such approaches, however, do not generally account for uncertainty in the ordering. Nor do these approaches typically provide an easily interpretable relationship between the inferred pseudotime and chronological time. The latter effect is compounded when inferring pseudotime for datasets with multiple branches: if one branch has fewer observations, or else a period of quiescence, the branch will often be artificially truncated compared to others. This truncation can be undesirable when investigating complex developmental programs, where signalling from adjacent tissues can influence cell fate decisions, and is necessary to place branches on consistent pseudotime frames. Finally, due to decreased costs, scRNA-seq datasets are now routinely generated over more finely resolved time series, including early stages of development (Yan et al. 2013; Guo et al. 2015; Petropoulos et al. 2016; Borensztein et al. 2017; Huang et al. 2017) and PGC development (Guo et al. 2015; Li et al. 2017). The individual populations of cells associated with different stages may still reflect a continuum of states, with some degree of overlap between stages; nevertheless, the capture time should be highly informative about the ordering of cells. To address these various limitations, further methodological developments are required to extract the maximal amount of information from such data.

Nonparametric Bayesian approaches based on Gaussian process (GP) regression (Rasmussen and Williams 2006), would therefore be advantageous, providing a flexible framework that better reflects the underlying generative models, and are able to quantify a broad range of dynamic behaviour with minimal need for parametrization. GP-based models provide robust treatments of uncertainty, and can gracefully cope with missing observations, for example the absence of intermediate cell types. Recent developments by Reid and Wernisch (2016) have demonstrated how cells can be ordered along a single developmental axis using Gaussian process latent variable models (GPLVMs) by placing informative priors over pseudotimes. This can be thought of as a dimensionality reduction technique with one latent dimension, pseudotime, that is *a priori* centered on the capture time. Until recently, however, GPs could not be used to model branching processes, and these methods were thus inapplicable to developmental trajectories with more than one branch. Recent developments by Yang et al. (2016) and Penfold *et al.*, (2018, submitted), have developed explicit branching Gaussian process frameworks (B-RGPs). Here we use a branching covariance function within a GPLVM (B-RGPLVM) to dissect the transcriptional program and signalling pathways implicated in human PGC competence and specification using scRNA-seq data from PGCs, somatic tissues, preimplantation 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. Analysis of the earliest branching events highlighted possible 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.

2 Methods

The ability to leverage capture time and other prior information, as well as the ability to quantify uncertainty in pseudotemporal ordering is particularly important given the inherent limitations of pseudotime approaches (see e.g., discussion in Campbell and Yau (2015); Weinreb et al. (2017)). Our model assumes we have M cells that have been assayed at one of $T < M$ distinct known capture times $\boldsymbol{t}_c = (t_1, \dots, t_M)$. The aim is to infer a corresponding set of pseudotimes, $\boldsymbol{t}_p = (\tau_1, \dots, \tau_M)$, such that the gene expression profiles are smooth and reflect the general developmental trajectory of the process.

Within the GPLVM model of Reid and Wernisch (2016), the pseudotimes, $\mathbf{t_p}$ are assigned a prior that is conditioned on the capture times, $\mathbf{t}_c$, that is Gaussian distributed: $\mathbf{t}_p|\mathbf{t}_c \sim \mathcal{N}(\mathbf{t}_c, \sigma^2\mathbb{I})$. The GPLVM model assigns an independent Gaussian process prior to the expression profile f_i of each gene i over pseudotime, $f_i \sim \mathrm{GP}(\mu_i(t_p), k_i(t_p, t_p{}'))$, where $\mu_i(t_p)$ denotes the mean function and $k_i(t_p, t_p{}')$ the covariance function. Different choices of covariance function encode different prior distributions over the expression profiles of the individual genes. Reid and Wernsich (2016) assume a Matérn covariance function:

$$k_i(t_p, t_p') = \psi_i\,(1 + \sqrt{3}\,|t_p - t_p'|/\ell)\exp(\sqrt{3}\,|t_p - t_p'|/\ell)\,.$$

where ℓ is a global length-scale and ψ_i are gene-specific scaling factors. The expression data for each gene i over all cells are modelled as noisy versions of their underlying expression profiles, $\mathbf{y}_i \sim \mathcal{N}(f_i(\mathbf{t}_p), \omega_i^2 \mathbb{I})$, where the ω_i are gene-specific noise levels. As the Gaussian process prior is conjugate to this likelihood, f_i can directly marginalized:

$$\int P(\boldsymbol{y}_i|\boldsymbol{f}_i, \boldsymbol{t}_p) P(\boldsymbol{f}_i|\boldsymbol{t}_p) d\boldsymbol{f}_i = \boldsymbol{N}(\boldsymbol{y}_i|\boldsymbol{\mu}_{\boldsymbol{t}_p}, \boldsymbol{K}_{\boldsymbol{t}_p,\boldsymbol{t}_p} + \omega_i^2 \mathbb{I}),$$

where $\boldsymbol{\mu}_{\boldsymbol{t}_p}$ represents a vector of the mean function evaluated at the pseudotimes times $\boldsymbol{t}_p$, and $\boldsymbol{K}_{\boldsymbol{t}_p,\boldsymbol{t}_p}$ the covariance matrix. The composite likelihood for pseudotimes can be written as:

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

Until recently no explicit GP treatment for branching processes existed, and earlier works were thus limited to single developmental trajectories (Reid and Wernisch 2016) or required pseudotemporal ordering to be run independently on the individual branches. Recent studies by Yang *et al.* (2016) and subsequent work by Penfold *et al.*, (2018, submitted) have shown how compositions of covariance functions can be used to explicitly define a branching Gaussian process. Specifically, we consider a two-branch covariance function, which can be constructed via the following composition of covariance functions:

$$K_{2-branch}(t_p, t_p', z, z') = k_1(t_p, t_p') + CP_{t_{b,s}}(K_0, k_2(t_p, t_p'))\delta_{z,2}\delta_{z',2} + \beta\delta_{t,t'}\delta_{z,z'},$$

where $K_0 = K_0(t, t')$ denotes a zero-kernel, and $CP_{t_b}(k_1, k_2)$ denotes a change-point kernel (Lloyd et al. 2014) with branch time $t_b$and slope parameter s. Here z and z' are branch labels that correspond to t_p, t_p'; $\delta_{z,2}$ is the Dirac delta function that is 1 if and only if $z = 2$. Finally, k_1 is the kernel for branch 1 and k_2 is the kernel that models the deviation of branch 2 from branch 1.

or cells sampled from a bifurcating developmental process with capture times, $\boldsymbol{t}_c = (t_1, \dots, t_M)$, and branch-labels, $\boldsymbol{z}_c = (z_1, \dots, z_M)$, each gene can thus be assigned an independent B-RGP prior over the pseudotimes, and the composite likelihood is:

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

where $\boldsymbol{y}_i$is the vetctor of gene expression for the *i*th gene. We may easily sample from this using Markov chain Monte Carlo.

3 Results

To assess the ability of B-RGPLVMs to pseudotemporally order data over a branching process, we first 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) an infection time-series showing changes in gene expression in *Arabidopsis* at two hour intervals over a 48h period after inoculation with *B. cinerea*; and (ii) a time-matched control time series. In general, we note that B-RGPLVMs outperforms existing approaches for pseudotime ordering (compare Supplementary Figures 1 and 2; for full details see Supplementary Section 1). Accuracy of the models increases as the number of marker genes increases. For the case $N = 40$, 4 out of 5 replicates were nearly perfectly ordered for both the infected and control time series, and no obvious improvements seen for $N = 80$. Consequently, when pseudotemporally ordering data, we used >40 genes.

3.1 Transcriptional branching and recombination in PGCs

Early human PGC development remains underexplored, since the use of human embryos at the time of PGC specification is limited by ethical and practical considerations. 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). One such model relies on derivation of PGC-like cells (PGCLCs) from human pluripotent stem cells (hPSCs) *in vitro* (Irie et al. 2015; Sasaki et al. 2015b); for this, conventional hPSCs are first placed in appropriate media to gain germline competence and then exposed to BMP signalling to form PGCLCs. An added advantage of this system is the ability to use genetic and pharmacological tools to address the roles of particular regulators in human PGC fate. Unbiased approaches involve genetic and chemical screens, which tend to be costly and time-consuming. 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.

Table 1. Number of cells for each cell type

Stage	*N*	Source
ESCs	34	Yan et al. (2013)
Pre-implantation	90	Yan et al. (2013)
PGCs	242	Guo et al. (2015)
Soma	86	Guo et al. (2015)

Here we have used B-RGPLVMs to infer the transcriptional branching dynamics of early human PGCs to identify the earliest regulators of PGC fate. For this we used scRNA-seq data from hPGCs and age-matched neighbouring somatic tissues from weeks 4 through to 19 (Guo et al. (2015); GEO GSE63818), as well as pre-implantation embryos (oocytes, zygotes, two-cell, four-cell, morula and blastocyst-stage cells), and conventional hESCs at passages 0 and 10 (Yan et al. (2013); GEO GSE36552). A summary of the various cell types is included in Table 1. Together, these datasets allow investigation of the dynamics of decision making in early human embryos, as cells undergo specification towards the germline, and provide a rich testbed for addressing the differences between the *in vivo* dynamics of PGC development, and the inferred dynamics of pseudotemporally ordered hESCs recombining with Wk4 PGCs. Whilst there is currently no single-cell transcriptomic data available for PGC-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.

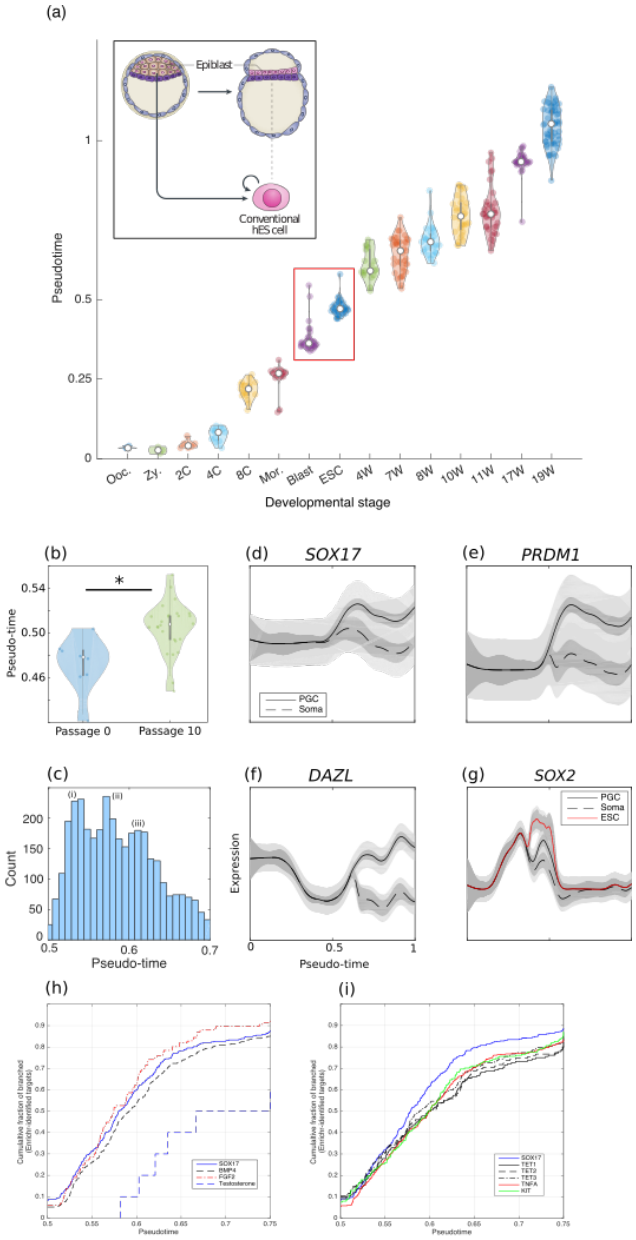


Figure 1: (a) Inferred pseudo-temporal ordering of individual cells over a two-component branching component correctly identifies the developmental ordering of blastocyst stage cells and ESCs, suggesting hESCs are developmentally more advanced than blastocyst stage. **(b)** ESCs at passage 10 appear 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, branches 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.

3.1.1. Benchmarking of B-RGPLVMs

To separate out the developmental trajectories of PGCs from those of gonadal somatic cells we 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 centered on the cells' developmental stages (Reid and Wernisch 2016). Preliminary benchmarking of branch-recombinant GPLVMs (B-RGPLVM) using other datasets suggested that cells could be accurately (pseudo-temporally) ordered along a branching process using around 40 genes, and we therefore used the expression of 44 marker genes identified in Irie et al. (2015), which included PGC, pluripotency, mesoderm, and endoderm markers.

For comparison purposes we also pseudotemporally ordered cells using previously established methods, including Monocle2 (Qiu et al. 2017), TSCAN (Ji and Ji 2016). We also compared to a variety of other approaches including Wishbone (Setty et al. 2016), SCUBA (Marco et al. 2014), and SLICER (Welch, Hartemink, and Prins 2016), and evaluated the accuracy of these methods by comparing the correlation between the inferred pseudotime, and the capture time (Supplementary Figure 3). Overall, the B-RGPLVM offered the best performance, with Pearson' correlation of 0.972 for the PGC branch and 0.966 for the soma branch. Other approaches also performed well. For example, Monocle2 correlated well when the PGC branch and soma branches were considered separately, with correlation 0.894 and 0.922 for the PGC and soma branches respectively, although accuracy was degraded when both branches were considered together. TSCAN had a Pearson's correlation of 0.8534 and 0.918. All other approaches showed a general ability to separate out pre-implantation cells from PGC or soma, and showed correlation >0.7 for both branches. A major difference between previous approaches and our B-RGPLVM is the ability to place the different branches on a similar time frame. Only B-RGPLVMs and Monocle2 appeared to show any degree of alignment between the branches, with B-RGPLVM performing better overall (Supplementary Figure 3). Benchmarking against most widely used methods revealed that B-RGPLVM is better suited for pseudotime ordering of cells along multiple branches and places the cells closer to their real capture time.

We further evaluated the accuracy of our B-RGPLVM using unlabelled data, assuming that the branch-label and capture-times for hESC and blastocyst cells were unknown variables to be inferred within the algo-

rithm (see Supplementary Materials Section 2.3). Our inference correctly placed the unlabelled late-blastocyst cells between morula-stage and week 4 (Wk4) cells (Figure 1(a)). Furthermore, whilst hESCs are derived from the ICM of blastocysts, our approach placed them between day 6 blastocysts and Wk4 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 conventional hESCs sharing characteristics with the post-implantation epiblast (Tang et al. (2016); Figure 1(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 1(b)), potentially reflecting culture-induced adaptations.

3.2.2. Inferring branching structure on a genome scale

Once we had established a point estimate for the pseudotemporal order of cells using the chosen set of 44 marker genes, we used B-RGPs to identify complex patterns of branching present on a genome scale. Note that to do so, we fixed pseudotimes according to the ordering at step 30,000 in our earlier pseudotime analysis, and the results of this section do not take advantage of the uncertainty over the pseudotimes. For each gene, performing B-RGP regression (Penfold et al., 2018, submitted) by fitting one of three models and using the BIC to select the optimal structure. The three models are as follows: (i) somatic cells branched from the base process (hPGCs), with hESCs and PGCs following an identical process (see Figure 1(d, e, f)); (ii) both soma cells and hESCs branched from the main process, with hESCs later recombining towards Wk4 PGCs (Figure 1(g)); and (iii) all processes were identically distributed (no branching). A summary spreadsheet is available in Supplementary Table 4, and more details about the models found in Supplementary Section 3. The frequency of the three groups is indicated in Supplementary Figure 4(a), which shows that most genes were not differentially expressed i.e., non-branching. A histogram of the timing of branching is indicated in Supplementary Figure 4(b, c), and shows that of the genes that diverge between PGCs and soma, most do so between blastocyst stage and week 4, in line with the reported time of PGC specification (De Felici 2013).

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 5 (see also associated Supplementary Table 1). Our analysis revealed several early terms related to BMP and WNT signalling, with subsequent terms associated with epigenetic reprogramming, proliferation and cell migration, and later terms relating to testosterone signalling, meiosis, and gametogenesis which, together, reflects 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 1(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 1(d, e, f)). 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 showed that common genes included the PGC regulator *PRDM1*, known to act downstream of *SOX17* (Irie et al. 2015). Crucially, the overlapping targets of *SOX17* were, themselves, shown to branch earlier than average branching times in the dataset as a whole ($p<1x10^{-7}$, Kolmogorov–Smirnov test), and earlier then testosterone signalling, involved in later male germ cell development (Figure 1(h)). Together, these results highlight the crucial early role for *SOX17* in hPGC fate decision.

To reveal 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 used a Kolmogorov-Smirnov (KS) test to identify genes branching significantly earlier than average branching times in the dataset. 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 Enrichr 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, we envisage that in this way it may be possible to identify signalling molecules secreted from surrounding tissues and therefore not branching in the examined transciptome itself.

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 1(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; Oosterhuis and Looijenga 2005). 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 3), suggesting a possible functional role. 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* ($p<1x10^{-6}$, Enrichr; 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 line with existing experimental evidence (Tang et al. 2015). Amongst the related enriched GO terms, we noted several related to cAMP activity, consistent with recent studies in mice identifying the roles of cAMP signalling in mouse PGC proliferation and epigenetic reprogramming (Ohta et al. 2017).

In addition to picking up known PGC regulators, we were also able to identify other pathways potentially involved in human PGC fate, including a putative role for tumour necrosis factor (TNF) signalling in PGC development, based on the enrichment of 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 to perturbations of *TET1*, *TET2*, and *TET3* (Figure 1(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 we cannot exclude 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 the nerve fibres to 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 pseudotemporally ordered hESCs. In particular, the ability to derive human PGC-like cells (PGCLCs) from hESCs *in vitro* prompted us to regard hESC to PGC transition as a recombination process. The comparison of PGCLC transcriptome with that of *in vivo* PGCs suggests that PGCLCs closely resemble pre-migratory PGCs (Irie et al. 2015). We therefore considered the recombination to occur between hESC and the earliest available PGC dataset (Wk4). Thus, the comparison of hESC and hPGC dynamics represents a branch-recombinant process, whereby specified hPGCs and hESCs branch from a common precursor (blastocyst), with hESCs dynamics allowed to recombine with W4 hPGCs upon exposure to appropriate stimuli.

This analysis identified 1,331 branching genes, 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 4(c). GO analysis of these groups identified a number of enriched terms, including those relating to 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 2).

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 overexpression 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), potentially highlighting an 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 is added as to the *in vitro* culture medium used to derive PGCLCs from hESCs (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. Indeed SCF is 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 *SOX2* knockout ($p<5x10^{-23}$). *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 identifies key regulators of competence for PGC fate

Human ESCs 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 high temporal resolution snapshots of the underlying transcriptional dynamics, 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 1(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 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 2). 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).

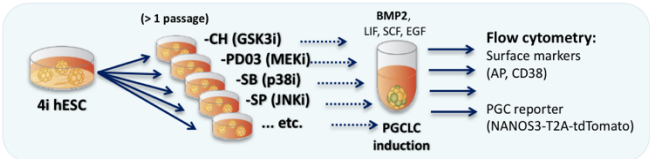


Figure 2. Scheme of the experimental workflow used to test the roles of individual components of the 4i hESC medium to sustain PGCLC competence. hESCs cultured in 4i medium were transferred to media lacking one of the kinase inhibitors for at least one passage. These cells were then subjected to standard 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.

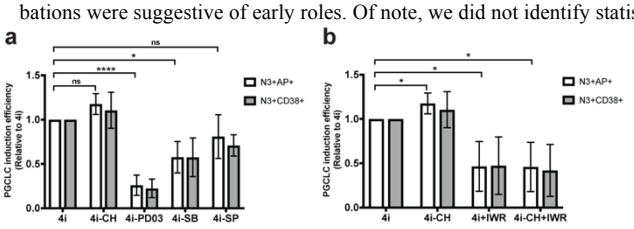


Figure 3. Quantification of PGCLC induction efficiency from hESCs grown in indicated conditions relative to 4i hESCs. PGCLCs induction efficiency was defined as the percentage of live NANOS3-tdTomato/AP (N3+AP+) or NANOS3-tdTomato/CD38-double positive cells. Data are shown as mean ± SD of 2 or 4 independent 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.

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 3(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 3(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 7).

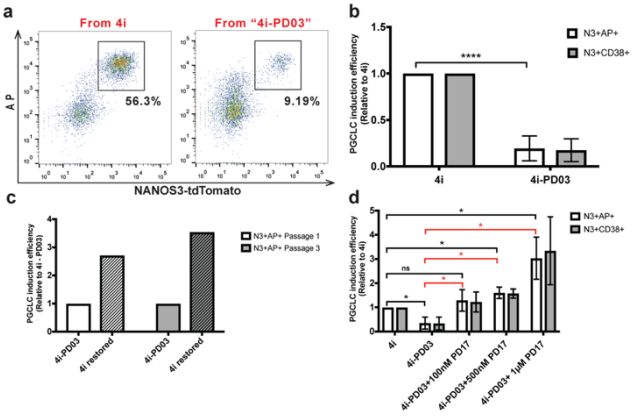


Figure 4. FGFR-MEK signalling is a negative regulator of human PGCLC competence. **(a)** MEK inhibitor withdrawal from 4i medium decreases PGCLC competence of hESCs. Representative flow cytometry plots of EBs derived from 4i and "4i-PD03" hESCs. **(b)** Quantification of PGCLC induction efficiency from "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.

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 4(b)), although unlike "4i-CH" or "4i-SB", "4i-PD03" cells could be maintained for many passages (>20). Importantly, the defect of "4i-PD03" hESCs to differentiate to PGCLCs could be rescued (within one passage) if PD03 was reintroduced (Figure 4(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 a collapsed after 2-3 passages. Importantly, FGFRi showed a dose-dependent positive effect on competence (Figure 4(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.

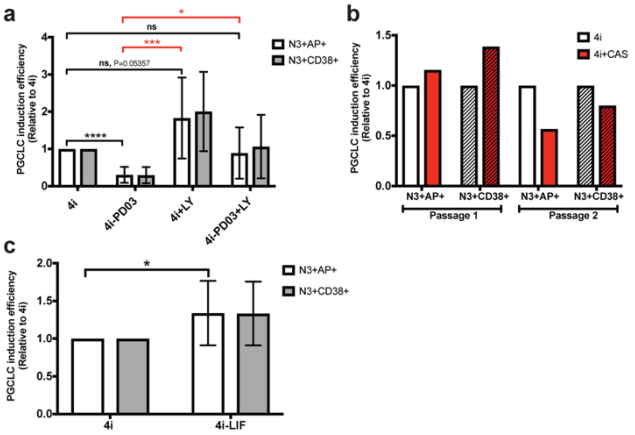


Figure 5. PI3K and LIF/JAK/STAT pathways negatively regulate PGCLC competence. **(a)** PI3K inhibition promotes PGCLC competence. Quantification of PGCLC induction efficiency from cells grown in indicated conditions relative to 4i hESCs. Data are shown as mean ± SD of 10 independent experiments. * $p \leq 0.05$, *** $p \leq 0.001$, **** $p \leq 0.0001$, 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).

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 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 hESC to hPGC 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 com-

plete 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 5(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 5(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 5(b)). This is in line with the enhanced competence of hESCs cultured in 4i without LIF, an agonist of JAK-STAT (Figure 5(c)). Interestingly, differentiation of 4i+CAS" hESCs yielded more CD38^{+} cells (Figure 5(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 5(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 6(a)). Intriguingly,

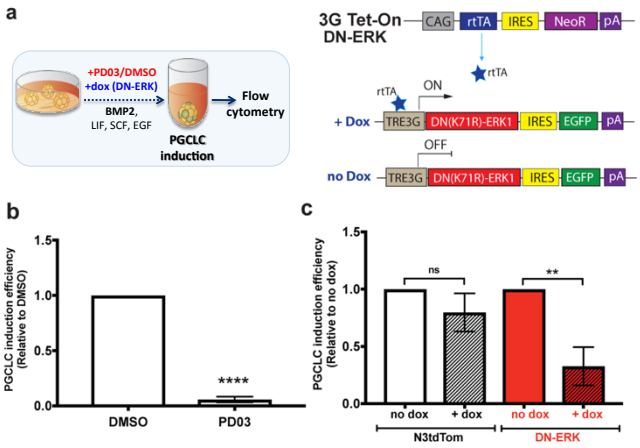


Figure 6. MEK-ERK inhibition during PGCLC induction interferes with PGCLC specification. **(a)** Workflow of the experiments and the scheme of the construct for DN-ERK transgene expression. Either PD03 or DMSO (vehicle) was added to the differentiation medium. Alternatively, to induce DN-ERK, PGCLC medium 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.

both perturbations reduced PGCLC numbers compared to controls (Figure 6(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 8). Altogether, this shows that the timing of FGFR signalling is crucial for context-specific cell fate decisions and our methodology allows predicting relevant changes in signalling and gene expression in a time-resolved manner.

4 Discussion

Transcriptional branching and recombination are frequently encountered in developmental biology and upon system perturbations. Differentiation of stem and progenitor cells to diverse 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, during dedifferentiation of cancer cells (Gupta et al. 2011; Friedmann-Morvinski and Verma 2014), or 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). 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 investigation of decision-making processes increasingly utilises single cell transcriptomic datasets. Typically, this requires a preliminary dimensionality reduction, with individual cells ordered over the reduced dimensional space, ideally in a manner that reflects developmental progression. Such approaches, however, have primarily been designed with heterogeneous populations in mind, with samples taken from single or lowly resolved time points, rather than populations taken from reasonably resolved developmental time series. Furthermore, these methods do not typically deal with uncertainty in the ordering of cells, and provide no easily interpretable relationship between pseudotime and chronological time.

Here by using an explicit statistical model for branching within a GPLVM, we have a statistical model for pseudotime that is not only reflective of the generative model of cell fate decisions, but is also apt to capture important information about the system such as capture time. Our approach is specifically designed with highly resolved single-cell RNA-seq time series in mind, and is shown to provide more accurate orering of cells compared to other approaches such as Monocle2 and TSCAN. Nonetheless, these gains require significantly more compute power, and will not, generally be as applicable for datasets with few time observations.

Having established the advantages of B-RGPLVMs we used our approach 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 correctly identified 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 lineage. Interestingly, mouse PGCs do not require *Sox17* and instead express *Sox2*. Many more unexplored differences exist between mouse and human PGC transcriptomes (Tang et al. 2015) and 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 cascades 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. Indeed, 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 in a principled way. 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 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.

Concurrent with our study, Boukouvalas, Hensman, and Rattray (2017) have built on the earlier work of Yang et al. (2016), to develop Branching Gaussian processes, a probabilistic framework for the inference of bifurcations in single cell transcriptional datasets. Unlike our approach, which utilised MCMC to deal with unlabelled datasets, these branching Gaussian processes focussed on efficient, scalable (variational) approximations for a two-branch system implemented using GPflow (Matthews et al. 2016), and were therefore similarly applicable even when cells lacked explicit branch labels. By identifying early branching events, they found known key regulators involved in the differentiation of haematopoietic stem cells into myeloid or erythroid precursors. Together with our own work, this demonstrates the usefulness of GP-based approaches for investigating complex single cell transcriptomic datasets.

In general, the use of B-RGPLVMs 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 6). Whereas 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, or at 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 relationship between transcriptional programs, the epigenetic state, and cell fate decisions.

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 useful advice, 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 Scholarship and Cambridge International Trust Scholarship. MAS is supported by HFSP and a Wellcome Trust Senior Investigator Award. JR and LW are funded by the UK Medical Research Council (Grant Ref MC_U105260799). YH is supported by a studentship from the James Baird Fund, University of Cambridge. MG acknowledges funding from 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.

Conflict of Interest: none declared.

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.

Bendall, S. C., K. L. Davis, A. D. Amir el, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er. 2014. 'Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development', *Cell*, 157: 714-25.

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.

Boukouvalas, Alexis, James Hensman, and Magnus Rattray. 2017. 'BGP: Branched Gaussian processes for identifying gene-specic branching dynamics in single cell data', *bioRxiv*, 10.1101/166868.

Campbell, K., and C. Yau. 2015. 'Bayesian Gaussian Process Latent Variable Models for pseudotime inference in single-cell RNA-seq data', *bioRxiv*: 10.1101/026872.

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.

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.

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: e22345.

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.

Ji, Z., and H. Ji. 2016. 'TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis', *Nucleic Acids Res*, 44: e117.

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.

Li, L., J. Dong, L. Yan, J. Yong, X. Liu, Y. Hu, X. Fan, X. Wu, H. Guo, X. Wang, X. Zhu, R. Li, J. Yan, Y. Wei, Y. Zhao, W. Wang, Y. Ren, P. Yuan, Z. Yan, B. Hu, F. Guo, L. Wen, F. Tang, and J. Qiao. 2017. 'Single-Cell RNA-Seq Analysis Maps Development of Human Germline Cells and Gonadal Niche Interactions', *Cell Stem Cell*, 20: 858-73 e4.

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.

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*.

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.

Marco, E., R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G. C. Yuan. 2014. 'Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape', *Proc Natl Acad Sci U S A*, 111: E5643-50.

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.

Matthews, Alexander G. de G., Mark van der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo Leon-Villagra, Zoubin Ghahramani, and James Hensman. 2016. 'GPflow: A Gaussian process library using TensorFlow', *arXiv*: 1610.08733.

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.

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.

Ohta, H., K. Kurimoto, I. Okamoto, T. Nakamura, Y. Yabuta, H. Miyauchi, T. Yamamoto, Y. Okuno, M. Hagiwara, K. Shirane, H. Sasaki, and M. Saitou. 2017. 'In vitro expansion of mouse primordial germ cell-like cells recapitulates an epigenetic blank slate', *EMBO J*.

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

Oosterhuis, J. W., and L. H. Looijenga. 2005. 'Testicular germ-cell tumours in a broader perspective', *Nat Rev Cancer*, 5: 210-22.

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.

Poincaré, H. 1885. 'Sur l'équilibre d'une masse fluide animée d'un mouvement de rotation', *Acta mathematica*, 7: 259-380.

Qiu, X., Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell. 2017. 'Reversed graph embedding resolves complex single-cell trajectories', *Nat Methods*, 14: 979-82.

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. 2015a. 'Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells', *Cell Stem Cell*, 17: 178-94.

———. 2015b. 'Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells', *CELL STEM CELL*.

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.

Setty, M., M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er. 2016. 'Wishbone identifies bifurcating developmental trajectories from single-cell data', *Nat Biotechnol*, 34: 637-45.

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.

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.

Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. 2014. 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nat Biotechnol*, 32: 381-86.

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.

Weinreb, Caleb, Samuel Wolock, Betsabeh K. Tusi, Merav Socolovsky, and Allon M. Klein. 2017. 'Fundamental limits on dynamic inference from single cell snapshots', *bioRxiv*: 10.1101/170118.

Welch, J. D., A. J. Hartemink, and J. F. Prins. 2016. 'SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data', *Genome Biol*, 17: 106.

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.

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.

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

1 Benchmarking of branch-recombinant Gaussian process latent variable models (BR-GPLVMs)

To assess our ability to pseudo-temporally order data over a branching process using B-RGPLVMs, we first benchmark using existing 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-hour 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). As outlined by 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 B-RGPLVM. Here we selected $10, 20, 40$, and 80 genes at random from the set of 150 previously used by Reid and Wernisch (2016). By doing so, we aimed to identify how the accuracy of the model changed with respect to increased number of gene observations, and thus empirically identify the number of genes required for accurate pseudotime ordering. We performed 5 different randomly initialised runs, with 10,000 samples in the MCMC chain, discarding the first 5,000 steps for burn-in. Run time was in the region of 10 hours, although this could be decreased using MATALBs parallel processing toolbox. Note that for each independently initialised run, a separate random selection of genes was chosen to gauge variability in the inferred ordering with respect to different genes.

Since the time series were generated using bulk microarrays, with measurements based on populations of cells, we expected a smoothly varying process, and thus for the covariance function we used a two-component branching covariance function composed of squared-exponential covariance functions.

$$K(t,t',z,z') = k_{mock}(t,t') + CP_{t_b,s}(K_0, k_{treatment}(t,t'))\delta_{z,2}\delta_{z',2} + \beta\delta_{t,t'}\delta_{z,z'},$$

where $k_{mock}(t,t') = SE_{\theta_{mock}}(t,t') = \sigma^2_{mock}\exp((t-t')/2{l_{mock}}^2)$ and $k_{treatment}(t,t') = SE_{\theta_{treatment}}(t,t') = \sigma^2_{treatment}\exp((t-t')/2{l_{treatment}}^2)$. The hyperparameters in the model therefore corresponded to the length-scale and process variance of the base process (the control data), the length-scale and process variance of the perturbed process (the infected data), and the change point time (branch time) and branch rate for the change-point kernel. Hyperparameters were initially fitted using maximum likelihood estimates based on the low-resolution capture times, $\theta \leftarrow \arg_\theta \max \mathcal{L}(\theta | \boldsymbol{t}_p = \boldsymbol{t}_c, \boldsymbol{y})$. Length-scale, process-variance, and noise hyperparameters were then fixed for the remainder of the inference. For the change-point time hyperparameter, t_b, we assumed a univariate smoothed box prior distribution with linear decay in the log domain, $P(t_b|a,b) = \sigma(\eta(t_b - a))(1 - \sigma(\eta(t_b - b)))$, where $\sigma(z) = 1/(1 + \exp(-z))$, and $a = 0$, $b = 33$, and $\eta = 33$. Within the MCMC algorithm, we considered the following updates:

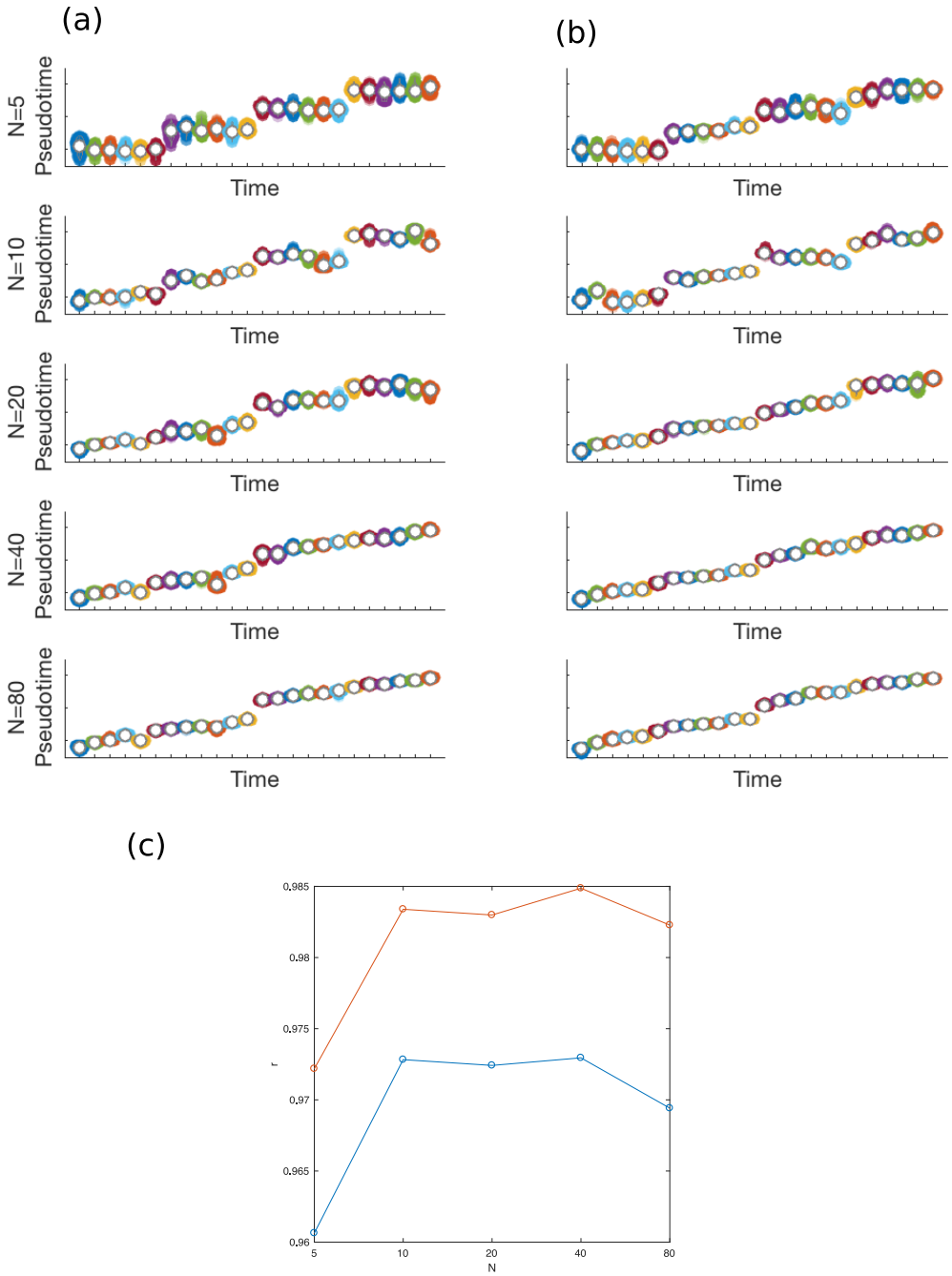
1. A Gaussian perturbation of a subset of pseudo-times.
2. A random swap of pseudotimes between adjacent data points.
3. A global swap of pseudotimes between two randomly selected data points.
4. Every 10 steps we sampled the change-point hyperparameters via HMC.

In Supplementary Figure 1, 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 $N \in (10,20,40,80)$. We note that, in general, the accuracy of the models appears to increase as the number of genes increases. For the case $N = 40$, 4 out of 5 replicates were nearly perfectly ordered for both the infected and control time series, with a mean correlation of >0.97 and >0.98 respectively over the five runs (for N=40). No obvious improvement was seen when increasing the number of marker genes to 80 (Supplementary Figure 1(c)). Consequently, when pseudo-temporally ordering data, we used >40 genes.

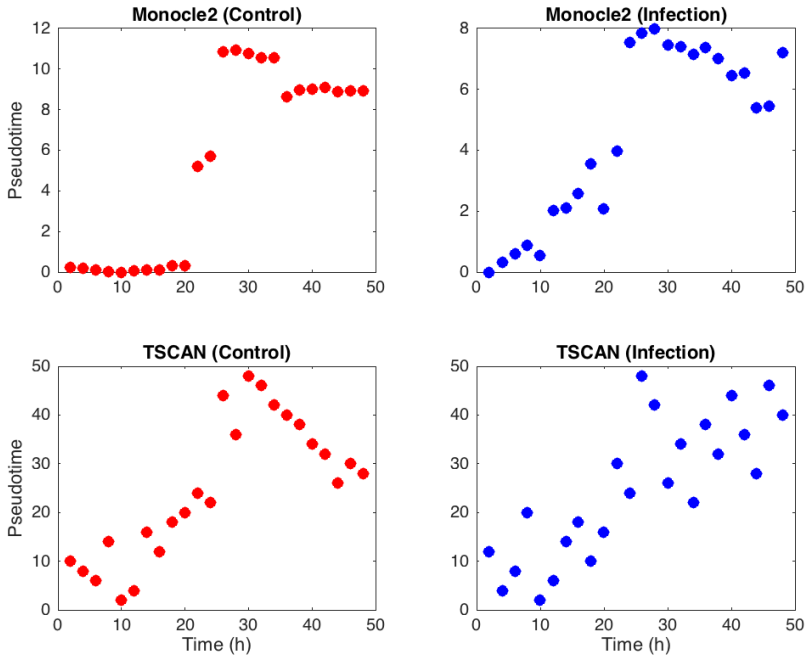
For comparison purposes, we also performed pseudotime using the full list of 150 marker genes previously used by Reid and Wernisch (2016) as taken from the main text of Windram et al. (2012). Here we used two approaches:

- Monocle2 (Qiu et al. 2017). We first ran Moncle2 using the combined control and treated datasets. Subsequently we ran it on the control and treated datasets separately. The later analysis showed higher correlation between inferred pseudotime and measurement time and are the results reported here.
- TSCAN (Ji and Ji 2016). We used the online TSCAN app (https://zhiji.shinyapps.io/TSCAN) on the control and treated datasets separately. As the data was already log normalised we applied no normalisation. We used PCA for dimensionality reduction, with number of components set using the 'Automatically select optimal dimension for PCA' option. Finally the number of clusters was selected with the 'Use optimal cluster number' option, and pseudotimes exported as csv files.

In supplementary Figure 2 we plot the measurement time versus inferred pseudotime for Moncle2 and TSCAN respectively for the control and infected datasets. For TSCAN we saw Pearson correlation of 0.74 and 0.81 respectively in the control and infected datasets, whilst for Monocle2 we found a correlation of 0.85 and 0.83 respectively. For the BRGPLVM we found a mean correlation of 0.97 and 0.99 respectively over five runs (for N=40).



Supplementary Figure 1: The pseudotemporal order (y-axis) is plotted against the true order (x-axis) for *Arabidopsis thaliana* transcriptional data for the control time-series (a) and *Botrytis cinerea* infected time series (b). (c) The mean correlation (over 5 random initialisations) between time and pseudotime as the number of marker genes increases.



Supplementary Figure 2: The pseudotemporal order (y-axis) is plotted against the true time (x-axis) for *Arabidopsis thaliana* transcriptional data using Monocle2 (Top) and TSCAN (Bottom) respectively.

2 PSEUDO-TEMPORAL ORDERING OVER DEVELOPMENTAL TRAJECTORIES

Single cells from pre-implantation embryos, PGCs, somatic cells and ESCs (Guo et al. 2015; Yan et al. 2013) were initially pseudotemporally ordered over a two-component branching process using 44 marker genes. For each of the 44 genes, the trajectory for specification of PGCs was chosen to represent the base process, due to this class having the most data points, with developmental trajectories for soma representing the branch process. Pre-blastocyst stage cells were randomly assigned to either branch with equal probability, whilst branch labels for blastocyst stage cells and ESCs were inferred within the algorithm. We performed five randomly initialised runs, using 30,000 steps in the MCMC chain, taking approximately 24 hours using a single CPU, although runtime could be decreased using the MATLAB parallel computing toolbox. The order of cells at step 30,000 appeared to show good overall correlation across these five runs, with mean correlation coefficient, $\langle R \rangle = 0.9768 \pm 0.003$.

For comparison purposes we also pseudotemporally ordered cells using a variety of other pseudotime methods, including Monocle2 (Qiu et al. 2017), TSCAN (Ji and Ji 2016), Wishbone (Setty et al. 2016), SLICER (Welch, Hartemink, and Prins 2016), and SCUBA (Marco et al. 2014). In all cases we

- Moncole2 (Qiu et al. 2017),. Here we ran Monocle2 several times. Initially we included all *in vivo* cells and attempted to capture the bifurcations between soma and PGCs. Within the algorithm, we used SOX17 and WT1 as marker genes for terminal fates of PGC and soma, respectively. In the second instance, we separated out the data, first combining pre-implantation data with PGCs, to infer the PGC trajectory, and then combining pre-implantation with soma. The greatest correlation between developmental stage and pseudotime was found for the second run i.e., running on the two branches separately.
- TSCAN (Ji and Ji 2016). To generate pseudotimes we ran TSCAN using the online app. PCA was used for dimensionality reduction, and clusters were selected using the optimal cluster

number option. Finally, we manually tuned the ordering of clusters and selected the combination that gave the greatest correlation between developmental stage and pseudotime.

- SCUBA and SLICER algorithms were run using default settings.
- Wishbone was run using the accompanying GUI, using default settings. Due to numerical issues, we ran with a reduced number of genes compared to that used for Monocle2, TSCAN, SCUBA and SLICER.

We evaluated the accuracy of the various approaches by comparing the correlation between the inferred pseudotime order and developmental stage (Supplementary Figure 2). All methods appeared able to broadly separate out the different cells types e.g., pre-implantation, PGCs and soma, with B-RGPLVM performing best at capturing the developmental trajectory, followed by Monocle2 and TSCAN.

B-RGPLVM
SLICER
SCUBA
WISHBONE
TSCAN
Monocle2
Pseudotime
ESC
Ooc.
Zy.
2C
4C
8C
Mor.
Blast
4W
7W
8W
10W
11W
17W
19W

Supplementary Figure 3: Comparison of pseudotime ordering using MONOCLE, TSCAN, Wishbone, SCUBA and SLICER versus B-RGPLVM. Here we indicate the inferred pseudotime (y-axis) versus the true developmental stage. All approaches appear to be able to generally separate out pre-implantation cells from either PGCs or soma, with B-RGPLVMs performing best. Together, these observations highlight the increased accuracy afforded when including prior information about capture time in B-RGPLVMs and a more realistic generative model in terms of an underlying branching process. Note that the distribution of cells from pre-implantation embryos will be identical when comparing within a method.

3 COMPLEX BRANCHING DURING EARLY EMBRYO DEVELOPMENT IN HUMANS

Once a point estimate for the pseudotemporal order of cells had been established, we looked for more complex branch-recombinant structures on a genome scale, by fixing pseudo-times according to the order at step 30,000 in our earlier pseudotime analysis. As the pseudotimes are fixed, this is simply a case of fitting GP model with different branch structures (B-RGP regression). 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. This assignment reflects our expectation that divergence between PGCs and soma occurs post blastocyst. 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:

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

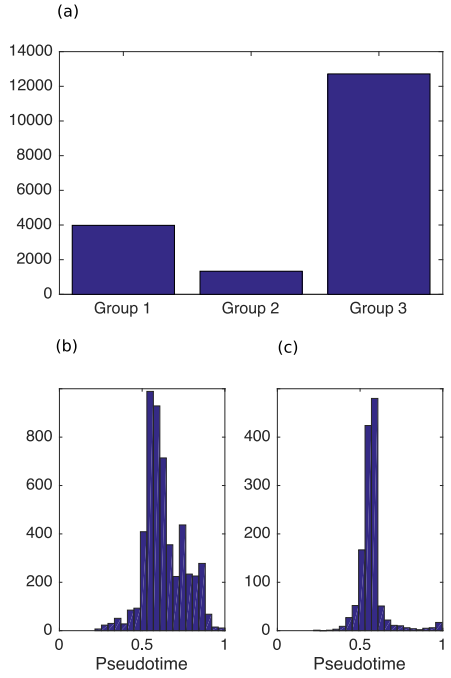
representing genes in which somatic cells diverged from PGCs, and where ESCs diverged from the *in vivo* dynamics of PGC specification before recombining at around week 4. For model 2, we assumed:

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

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

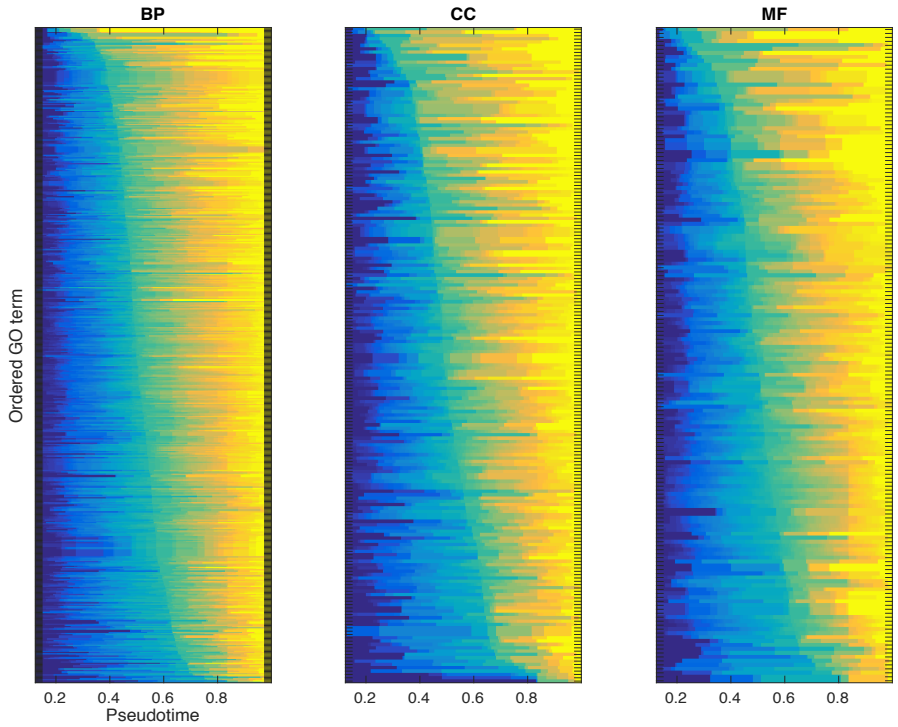
$$f_{PGC}(t), f_{soma}(t), f_{ESC}(t) \sim \mathcal{GP}\big(c, k_{PGC}(t,t')\big),$$

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 9(a), which shows that most genes were not differentially expressed i.e., non-branching. A histogram of the timing of branching is indicated in Supplementary Figure 9(b, c), and shows that of the genes that diverge between PGCs and soma, most do so between blastocyst stage and week 4, as expected.

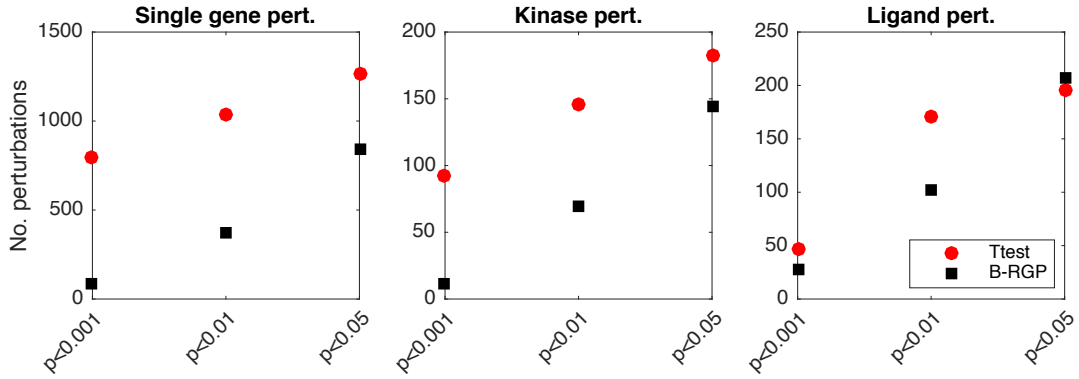


Supplementary Figure 4: (a) Frequency of the different groups. **(b)** Timing of branching between PGC and somatic cells. **(c)** Time of recombination between ESCs and inferred *in vitro* dynamics.

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 5: 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 6: 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 were 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 5) or modifications therefrom (Supplementary Table 6). 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 7) 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 Tables 1-4 are available as separate spreadsheets.

Supplementary Table 5. 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) 0.1 mg/ml (Streptomycin)	GIBCO
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 6. Modifications of the 4i medium used in the competence screen.

Medium	Recipe
4i-PD03	4i as in Supp. Table 5 without PD0325901
4i+PD17	4i as in Supp. Table 5 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 5 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 5 supplemented with 2.5 μM IWR-1 (Sigma)
4i-CH	4i as in Supp. Table 5 without CHIR99021
4i-CH+IWR	4i as in Supp. Table 5 without CHIR99021 supplemented with 2.5 μM IWR-1 (Sigma)
4i-SB	4i as in Supp. Table 5 without SB203580
4i-SP	4i as in Supp. Table 5 without SP600125
4i-LIF	4i as in Supp. Table 5 without human LIF
4i+JAKi	4i as in Supp. Table 5 supplemented with 1μM CAS457081-03-7

Supplementary Table 7. 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 8. After washing with FACS buffer, the cells were recorded on BD LSR Fortessa. Data were analysed using FlowJo (Tree Star).

Supplementary Table 8. 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 9). The $\Delta\Delta Ct$ method was used for quantification of gene expression. Three technical replicates were used for each biological replicate.

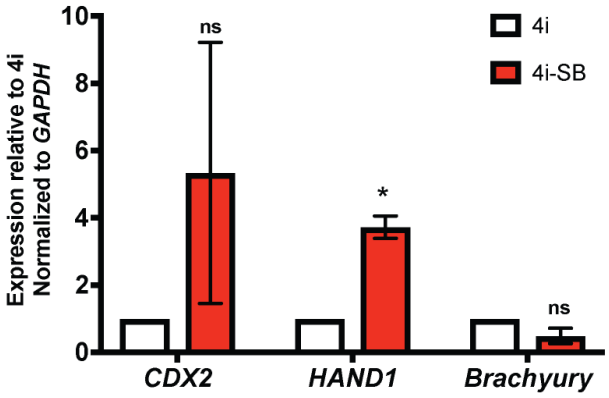
5.4. Dominant-negative ERK expression

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 6. 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.

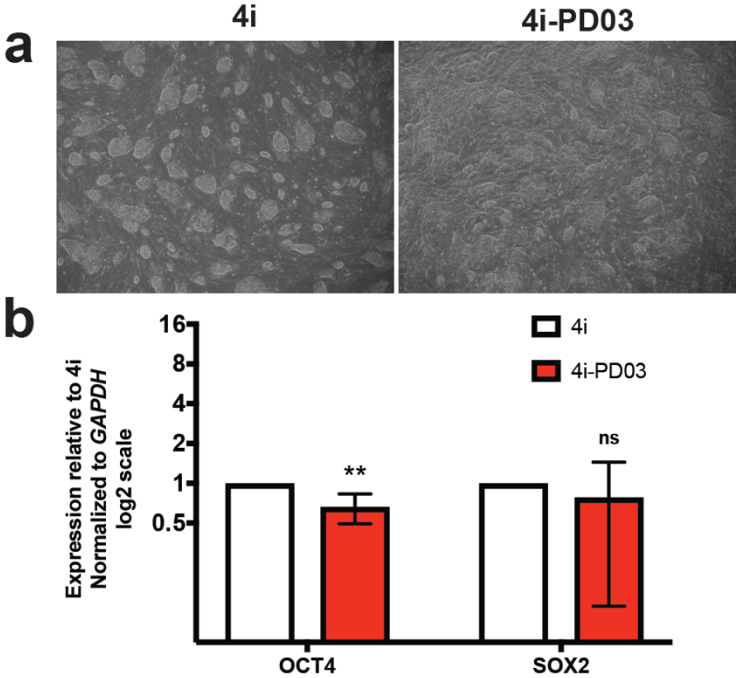
Supplementary Table 9. Oligos used in the study.

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

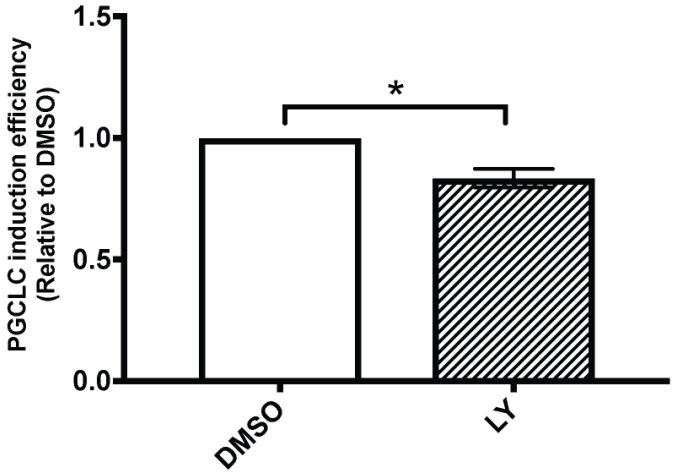
HAND1_h_F	qPCR	CGTGAGAGCAAGCGGAAAAG
HAND1_h_R	qPCR	AATCCTCTTCTCGACTGGGC
In-Fusion pPBTRE3G-hERK1_F	Cloning	ccctcgtaaagtcgaCGCCGCCACCatgGCGGCGGCGGCGGCTCAG
In-Fusion IRES2-hERK1_R	Cloning	GGGGAGGGAGAGGGGCTAGGGGGCCTCCAGCACTCCGGGCTG
In-Fusion pPBTRE3G-EGFP_R	Cloning	gggccccggtgtcgaCTTACTTGTACAGCTCGTCCATGCCG



Supplementary Figure 7. 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 8. 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 9. 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.

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.

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.

Ji, Z., and H. Ji. 2016. 'TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis', *Nucleic Acids Res*, 44: e117.

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*.

Marco, E., R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G. C. Yuan. 2014. 'Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape', *Proc Natl Acad Sci U S A*, 111: E5643-50.

Qiu, X., Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell. 2017. 'Reversed graph embedding resolves complex single-cell trajectories', *Nat Methods*, 14: 979-82.

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.

Setty, M., M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er. 2016. 'Wishbone identifies bifurcating developmental trajectories from single-cell data', *Nat Biotechnol*, 34: 637-45.

Welch, J. D., A. J. Hartemink, and J. F. Prins. 2016. 'SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data', *Genome Biol*, 17: 106.

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.

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.

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

Branch-recombinant Gaussian processes for analysis of perturbations in biological time series

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 Gurdon Institute, University of Cambridge, [2]Wellcome/MRC Cambridge Stem Cell Institute, University of Cambridge, [3]Physylogy, 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.

*To whom correspondence should be addressed. [§]These authors contributed equally.

Abstract

Motivation: A common class of behaviour encountered in the biological sciences involves branching or recombination. During branching, a statistical process bifurcates resulting in two or more potentially correlated processes that may undergo further branching; the contrary is true during recombination, where two or more statistical processes converge into one. The inference of branching and recombination from biological data remains a key challenge, with applications for differential expression analyses of multiple time series data, and dissecting the temporal progression of biological phenomenon. Gaussian processes (GPs) represents an ideal framework for such analysis, allowing for nonlinear regression that includes a rigorous treatment of uncertainty. Currently, however, GP models only exist for two-branch systems. Here we develop a general framework for the treatment of branching and recombination in the form of branch-recombinant Gaussian processes (B-RGPs).

Results: We first demonstrate the performance of B-RGPs compared to standard GP regression on a range of simulated datasets, and demonstrate robustness to model misspecification. B-RGPs are then used to investigate the branching patterns of *Arabidopsis thaliana* following inoculation with two strains of the hemibotrophic bacteria, *Pseudomonas syringae*. By grouping genes according to the number of branches, we could show enrichment for targets of pathogen protein effectors. Finally, we identify two early branching genes WRKY11 and WRKY17, and showed that groups of genes that branched at similar times to WRKY11/17 were enriched for W-box binding motifs, and overrepresented for genes differentially expressed in WRKY11/17 knockouts, suggesting that branch time could be useful for identifying direct and indirect binding targets of key transcription factors.

Availability: https://github.com/cap76/BranchingGPs

Contact : c.penfold@gurdon.cam.ac.uk, a.surani@gurdon.cam.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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 statistical process bifurcates at a specific time, leading to two potentially correlated processes that may, themselves, undergo further branching (Poincaré 1885). Reciprocal behaviour is encountered in recombination processes, where two or more statistical processes converge.

Such branching and recombination are frequently encountered in transcriptional time series datasets involving host-pathogen interactions. The initial response to infection is the activation of innate immunity, a highly-conserved response based upon perception of non-self. Subsequently, pathogens can deliver protein effectors which collectively suppress immunity, and later collaborate to reconfigure plant metabolism for pathogen nutrition. Thus, initially, the expression dynamics of key infection marker genes will be identically distributed in both infected and uninfected host cells. Expression patterns will begin to diverge as the host mounts immunity; in many cases, this innate immune response is suppressed by the pathogen, potentially driving expression levels of certain genes back to uninfected levels. Indeed nearly 50% of the transcriptome is observed to be differentially expressed during some plant infections (Windram et al. 2012; Lewis et al. 2015). More complex patterns of branching and recombination may exist in such datasets due to the ongoing evolutionary arms race between pathogens and their hosts (Jones and Dangl 2006; Boller and He 2009).

The inference of branching and recombination processes from systems level measurements, such as collections of microarray or RNA-sequencing data, remains a difficult challenge, partially due to datasets being noisy in nature, with (potentially) missing observations or uneven temporal sampling. The dynamic nature of different biological systems may also vary significantly, frustrating efforts to find a robust, broadly applicable approach to the inference of branching and recombination. Nonparametric Bayesian approaches to inference would therefore be advantageous, addressing these key issues. Gaussian processes represent a flexible nonparametric Bayesian approach to nonlinear regression able to gracefully cope with uncertainty, uneven sampling and a diverse range of dynamic behaviour (Rasmussen and Williams 2006). However, currently, Gaussian processes treatments for branching processes have only been developed for the two-branch case (Yang et al. 2016; Stegle et al. 2010). Here we develop an approach to inference for arbitrarily complex branching and recombination processes, in the form of branch-recombinant Gaussian processes (B-RGPs). In Section 2 we first introduce B-RGPs, highlighting their key limiting behaviour. In Section 3 we demonstrate the advantages of B-RGPs over GPs on a variety of simulated datasets, and in Section 4 we demonstrate the utility of our approach on genome-scale time-course microarray data by identifying transcriptional branching and recombination in *Arabidopsis thaliana* infected with bacterial pathogen *Pseudomonas syringae*. Finally, in Section 5, we discuss a variety of possible applications for B-RGPs and future avenues for research.

2 Methods

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). 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). 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 denote a vector of the observed expression levels for a given gene at times, $\boldsymbol{t}$. We can assign the unknown function a GP prior, denoted $f(t) \sim \mathcal{GP}(\mu(t), k(t,t'))$, 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 the more recent work of Yang et al. (2016), who demonstrate explicitly how a two-branch process can be encoded within a joint GP model using squared-exponential and Matérn covariance functions. To our knowledge, the generalisation of GPs to more than two branches has not been addressed to date. Furthermore, whilst the mixture model of Stegle et al. (2010) allows for recombination, it relied on sampling based inference, and no explicit closed-form solution to recombination has been outlined.

A useful extension to the GP framework is the multiple output hierarchical Gaussian process (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, $k_2(t,\boldsymbol{t}')$:

$$f_1(t) \sim \mathcal{GP}\big(0, k_1(t,t')\big),$$
$$f_2(t) \sim \mathcal{GP}\big(f_1(t), k_2(t,t')\big).$$

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 easily be encoded using HGPs. A class of branching behaviour can naturally be encoded within this HGP framework, 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 that ensures the two processes are identically distributed prior to an arbitrarily chosen time point, t_b:

$$f_1(t) \sim \mathcal{GP}\big(0, k_1(t,t')\big),$$
$$f_2(t) \sim \mathcal{GP}\big(f_1(t), CP_{t_b}(K_0, k_2(t,t'))\big).$$

where $K_0 = K_0(t,t')$ denotes a zero-kernel, and $CP_{t_b}(k_1,k_2)$ denotes a change-point kernel (Lloyd et al. 2014), defined as:

$$CP_{t_b}\big(k_1(t,t'), k_2(t,t')\big) = \sigma(t)k_1(t,t')\sigma(t') + \big(1-\sigma(t)\big)k_2(t,t')\big(1-\sigma(t')\big),$$

where $\big(1-\sigma(t)\big) = 1 + \tanh\left(\frac{t_b - t}{s}\right)/2$. Here we introduce two hyperparameters: t_b, which represents the branch time, and s, which controls how fast data belonging to the second branch transition from the basal process to the branch process. Each data point is assigned a branch label, $z \in [1,2]$, and observations are *a priori* Gaussian distributed: $\boldsymbol{y}_1, \boldsymbol{y}_2 |\, \boldsymbol{t}, \boldsymbol{z} \sim N(\boldsymbol{0}, K(\boldsymbol{t}, \boldsymbol{t}', \boldsymbol{z}, \boldsymbol{z}'))$, where:

$$K(t,t',z,z') = k_1(t,t') + CP_{t_b}(K_0, k_2(t,t'))\delta_{z,2}\delta_{z',2} + \beta\delta_{t,t'}\delta_{z,z'},$$

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 make a prediction $\boldsymbol{y}^*$ at a new input location, $(\boldsymbol{t}^*, \boldsymbol{z}^*)$, and analytically evaluate the marginal likelihood, allowing us to compare the goodness of fit between different branching processes.

Within our framework, we can allow further branches that independently diverge from the main branch, with an n-component system

defined by the following covariance function:

$$K(t,t',z,z') = K_{b_1}(t,t') + \sum_{i=2}^{n} CP_{t_i}\big(K_0, k_{b_i}(t,t')\big)\delta_{z,i}\delta_{z',i} + \beta\delta_{tz}\delta_{t'z'}.$$

Finally, each branch itself could give rise to further branching in a recurrent manner e.g., a basal (main branch) from which a secondary branch diverges, with a third branching from the second and so forth. For an n-component recurrent branching system we have:

$$K(t,t',z,z') = \begin{cases} k_1(t,t') + \beta\delta_{t,t'}\delta_{z,z'}, & \min(z,z') = 1 \\ k_1(t,t') + CP_{t_2}\big(K_0, k_2(t,t')\big) + \beta\delta_{t,t'}\delta_{z,z'}, & \min(z,z') = 2 \\ \vdots & \\ k_1(t,t') + \sum_{j=2}^{n-1} CP_{t_j}\left(K_0, k_j(t,t')\right) + \beta\delta_{t,t'}\delta_{z,z'}, & \min(z,z') = n-1 \\ k_1(t,t') + \sum_{j=2}^{n} CP_{t_j}\left(K_0, k_j(t,t')\right) + \beta\delta_{t,t'}\delta_{z,z'}, & z = n, z' = n. \end{cases}$$

When observation data for all branches are specified over identical time points, the covariance matrix can be expressed in a more compact notation:

$$K(\boldsymbol{t},\boldsymbol{t}',\boldsymbol{z},\boldsymbol{z}') = \sum_{j=1}^{n} \boldsymbol{A}^{(j)} \otimes \bar{k}_j(\boldsymbol{t},\boldsymbol{t}') + \boldsymbol{A}^{(1)} \otimes \beta\mathbb{I},$$

where $\otimes$ denotes the Kronecker product, and:

$$\bar{k}_j(\boldsymbol{t},\boldsymbol{t}') = \begin{cases} k_1(\boldsymbol{t},\boldsymbol{t}'), & j = 1 \\ CP_{t_j}\left(K_0, k_j(\boldsymbol{t},\boldsymbol{t}')\right), & \text{otherwise,} \end{cases}$$

where m represents the number of unique time points, $\mathbb{I}$ represents an $(m \times m)$ identity matrix, and $\boldsymbol{A}^{(j)} = \boldsymbol{u}\,\boldsymbol{u}^{\intercal}$, with $\boldsymbol{u}$ representing a column vector of length n, with ones in elements j through n and zeros everywhere else. Far more complex branching patterns can easily be built via the correct composition of independent and recurrent branching covariance functions.

As well as building branching structures of arbitrary complexity, we further note that the dynamic behaviour of the individual branches may themselves be arbitrarily complex, comprised of any linear combination of positive semi-definite kernels. In Figure 1(a) we indicate example behaviour of 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

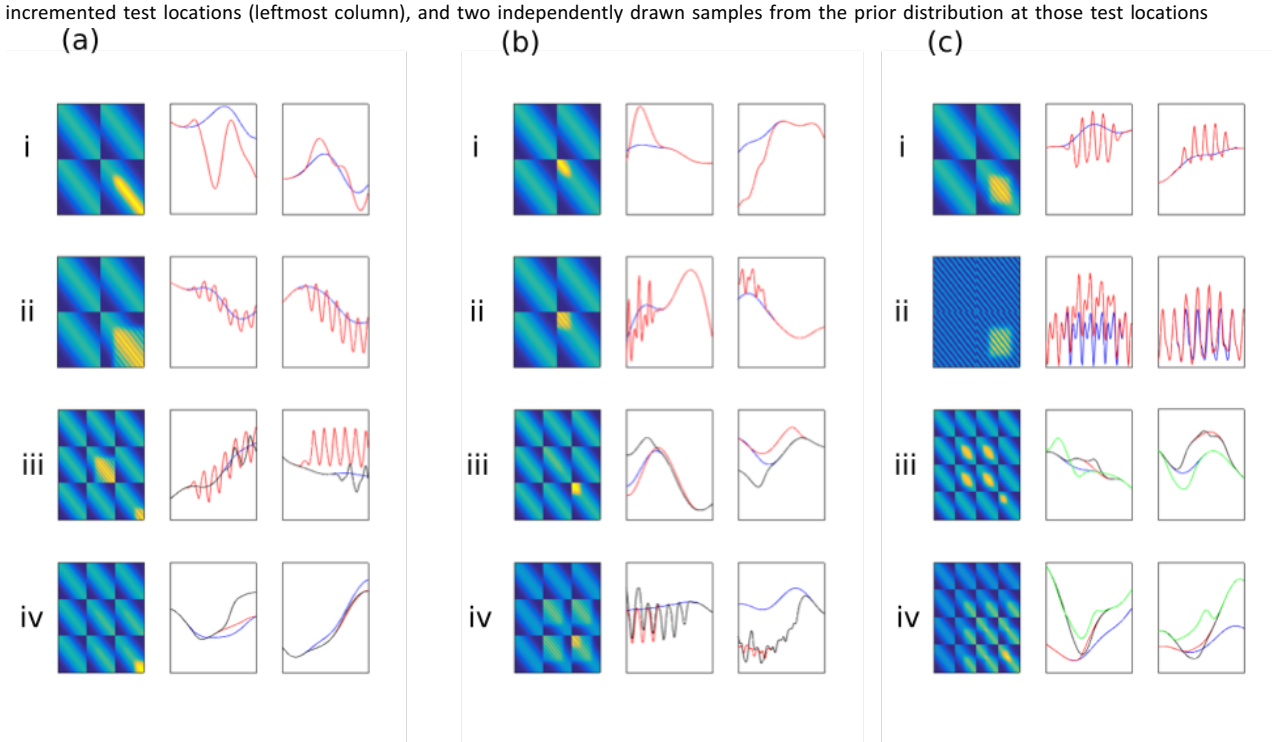


Figure 1: (a) GP with branching structure. *Here we indicate the prior covariance matrix evaluated at uniformly incremented input times (left) and two samples from the prior distribution (middle, right). (i) A smooth function branches from a smooth base function. (ii) A periodic function branches from 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, and a fourth function (blue) branches from the third function and recombines with the first two*

(middle and rightmost columns).

2.1 Recombinant Gaussian processes

Recombinant processes can be defined in a reciprocal fashion to branching processes. Notable examples might include the reprogramming of different terminally differentiated cell lineages to iPSCs (Gurdon 1962; Takahashi and Yamanaka 2006). We can describe a two-component system via the following composition of covariance functions:

$$K(t,t',z,z') = k_1(t,t') + CP_{r_b}(k_2(t,t'),K_0)\delta_{z,2}\delta_{z',2} + \beta\delta_{t,t'}\delta_{z,z'},$$

which encodes the main branch process, with a second (potentially correlated) process that recombines after time r_b. Multiple processes can again 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 Figure 1(b).

2.2 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. Such combinations of branching and recombination may be encountered during development when where there exists more than one route to a terminal cell fate, as may be the case in certain neuronal lineages (Zawadzka et al. 2010), as well as in certain diseases, such as during dedifferentiation of cancer cells (Friedmann-Morvinski and Verma 2014). An example two-component system can be encoded by the following covariance function:

$$K(t,t',z,z') = k_1(t,t') + CP_{r_2}\big(CP_{b_2}(K_0,k_2(t,t')),K_0\big)\delta_{z,2}\delta_{z',2} + \beta\delta_{t,t'}\delta_{z,z'}.$$

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 Figure 1(c).

2.3 Limiting behaviour

Depending upon the branch time and other hyperparameters in the change-point kernel, the behaviour of B-RGPs can naturally tend towards either an independent GP or a HGP. Specifically, for a branching GP, 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 low 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.

3 Results

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

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

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, IGPs were fitted to the two branches separately using squared-exponential covariance functions, with hyperparameters $\theta = [l_1, \sigma_1, l_2, \sigma_2]$ initialised as $\theta_i \sim U(0.1,1)$, and optimised using type II ML estimation. A JGP with squared-exponential covariance function was also fitted to the union of data, with hyperparameters $\theta = [l, \sigma]$ initialised and optimised as for the independent GPs. In Supplementary Figure 1 (top) we indicate an example posterior GP fit to the data using a BGP, IGPs and JGP respectively. In Supplementary Figure 1 (middle and bottom rows) we indicate the mean sum 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 Supplementary Figure 1(d) we plot the inferred branch time versus true branch time for 100 instances of a related branching process, in which the branch time is randomly (uniformly) distributed over the interval $t \in U(-8,8)$. We note that the correlation of R=0.99995 indicates good ability to infer branch times. To explore the ability to infer branch times for datasets with missing observations, we repeated this experiment, but excluded observations close to the true branch point, specifically removing any data points where $|t_{\hat{b}} - t_o| < 2$, where, $t_{\hat{b}}$ represents the true branch time, and t_o is the time of the data point. Even with missing observations centred at the true branch time, the inferred branch times were found to be highly correlated with the true branch time (R=0.9649; Supplementary Figure 1(e)).

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

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

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 $\boldsymbol{y}|\boldsymbol{t},\boldsymbol{z} \sim \mathcal{N}(f(\boldsymbol{t},\boldsymbol{z}),\sigma_n^2\mathbb{I})$, 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, IGPs and JGPS were also fitted, with hyperparameters initialised as previously described. Results are summarised in Supplementary Figure 2.

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. Nevertheless, 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 Supplementary Figure 3(a) we plot example fits to the function using a B-RGP, IGPs and a JGP. In Supplementary Figure 3 (b) we indicate the sum of squared errors for the B-RGP and a limiting case of the B-RGP (denoted with a *) 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 performed comparably to IGPs, although in a few instances the B-RGP appeared 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 was 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 branch-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, $f_{1,2}(t) \sim GP\left(0, K_0(t,t')\right)$. 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 Supplementary Figure 3 (c) we indicate example fits to the function using a B-RGP, IGPs and a JGP, whilst in Supplementary Figure 3 (d) 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.

Finally, in experiment 5 we performed inference on a four-branch system, in which we have one latent basal branch, from which two intermediate latent branches emerge. From each of the two intermediate branches we consider two observed branches. For comparison, we evaluate the sum squared error and log marginal likelihood for the B-RGP, IGPs and a JGP, with the results indicating B-RGPs provide better overall performance (Supplementary Figure 4).

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.

3.1. Inference for partially labelled datasets

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 cell types, including some that cannot be unambiguously assigned to a particular branch *a priori*. We can attempt to infer the branch labels, $\mathbf{z}$, using Markov chain Monte Carlo (MCMC). Here we assume partially labelled data, with a subset of branch labels know, and the remainder unknown, denoted $\mathbf{z} = [\mathbf{z}^{(\text{labelled})}, \mathbf{z}^{(\text{unlabelled})}]$. When branch labels are known, they can be fixed, whilst unknown branch labels are initialised stochastically, 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:

$$P(z_i = I | \boldsymbol{t}, \boldsymbol{z} \backslash z_i, \boldsymbol{y}, \theta) = \frac{P(\boldsymbol{y} | z_i = I, \boldsymbol{z} \backslash z_i, \boldsymbol{t}, \theta)}{\sum_{k=1}^{n} P(\boldsymbol{y} | z_i = k, \boldsymbol{z} \backslash z_i, \boldsymbol{t}, \theta)},$$

with hyperparameters updated conditional on all branch labels using hybrid Monte Carlo (HMC):

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

To test the accuracy of our B-RGPs on unlabelled or partially labelled data we generated observations from the simple branching process outlined in Supplementary Section 2. 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. A subset of data points, n, were assigned the correct branch label, where $n/N \in [0,0.1,0.25,0.5]$ and N indicates the total number of observations, with the remaining data points randomly assigned to either branch with equal probability and updated within the MCMC. Five randomly initiated runs were used, with 20,000 steps in the MCMC chain, and the first 5,000 discarded for burn-in. An example of the initial branch assignment is shown in Supplementary Figure 5(a), with red indicating data points initially assigned to branch 1, and blue assigned to branch 2. An example fit (and updated branch labels) is shown for step 20,000 in Supplementary Figure 5(b). The accuracy of classification is summarised using receiver operating characteristic (ROC) curves in Supplementary Figure 5(c, d). We note good overall ability to infer branch labels.

4 *Arabidopsis thaliana* transcriptional branching in response to *Pseudomonas syringae*

To evaluate the utility of B-RGPs on a genome scale applications, we used our framework to investigate transcriptional branching in model plant organism *Arabidopsis thaliana* in response to infection with hemibiotrophic bacterial pathogen *Pseudomonas syringae*.-Recent studies by Lewis *et al.*, 2015 (GEO GSE56094) have provided highly resolved transcriptional datasets for Arabidopsis following inoculation with disease-causing *Pseudomonas syringae* pv. tomato DC3000, and a disarmed mutant strain hrpA using bulk microarray measurements. Whilst both strains efficiently infect the plant cells, the DC3000 variant also delivers 28 effector proteins that subvert the plant's immune response; the disarmed hrpA mutant lacks the apparatus for effector delivery and thus elicits a classical immune response. Yang *et al.* (2016) developed a two-component branching GP to investigate transcriptional bifurcations between time series of hrpA- and DC3000-inoculated cells. Here we extend this analysis by simultaneously deciphering the branching structure that exists between all 3 time series (mock/control, virulent

(DC3000) and innate immune (hrpA) responses).

For each gene in the three datasets, we consider a number of possible branching structures: 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 Figure 2(a) in the main text.

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:

$$f_{mock}(t) \sim \mathcal{GP}\big(c, k_{mock}(t,t')\big),$$
$$f_{hrpA}(t) \sim \mathcal{GP}\left(f_{mock}(t), CP_{t_{b_1}}\left(K_0, k_{hrpA}(t,t')\right)\right),$$
$$f_{DC3000}(t) \sim \mathcal{GP}\left(f_{hrpA}(t), CP_{t_{b_2}}\left(K_0, k_{DC300}(t,t')\right)\right),$$

where observation data was 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:

$$f_{mock}(t) \sim \mathcal{GP}\big(c, k_{mock}(t,t')\big),$$
$$f_{hrpA}(t) \sim \mathcal{GP}\left(f_{mock}(t), CP_{t_{b_1}}\left(K_0, k_{hrpA}(t,t')\right)\right),$$
$$f_{DC3000}(t) \sim \mathcal{GP}\left(f_{mock}(t), CP_{t_{b_2}}\left(K_0, k_{DC300}(t,t')\right)\right).$$

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:

$$f_{mock,hrpA}(t) \sim \mathcal{GP}\left(c, k_{mock,hrpA}(t,t')\right),$$
$$f_{DC3000}(t) \sim \mathcal{GP}\left(f_{mock,hrpA}(t), CP_{t_{b_1}}\left(K_0, k_{DC3000}(t,t')\right)\right).$$

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:

$$f_{mock,DC3000}(t) \sim \mathcal{GP}\left(c, k_{mock,DC3000}(t,t')\right),$$
$$f_{hrpA}(t) \sim \mathcal{GP}\left(f_{mock,DC3000}(t), CP_{t_{b_1}}\left(K_0, k_{hrpA}(t,t')\right)\right).$$

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

$$f_{mock}(t) \sim \mathcal{GP}\big(c, k_{mock}(t,t')\big),$$
$$f_{hrpA,DC3000}(t) \sim \mathcal{GP}\left(f_{mock}(t), CP_{t_{b_1}}\left(K_0, k_{hrpA,DC3000}(t,t')\right)\right).$$

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 unbranched i.e., not differentially expressed. Because these datasets correspond to bulk observations from microarrays with well-defined measurement 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, and hyperparameters were optimised to their ML or 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 also Gamma distributed, $t_{b_1} \sim \Gamma(4,2)$, and the change-point transition rate was Gaussian distributed, $s \sim \mathcal{N}(0,0.5)$. 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; Stegle et al. (2010); Supplementary Figure 6(b)). Here the GP2S approach incorrectly identified 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 the perturbation times previously estimated in Yang et al. (2016) (Supplementary Figure 6(d)). The analysis in Yang et al. (2016) provide 90% confidence intervals for branch time 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 branch time in a subset of the observations may be unidentifiable.

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.

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, again using the BIC to distinguish between genes undergoing branching and recombination from those undergoing branching alone.

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

$$f_{DC3000}(t) \sim \mathcal{GP}\left(f_{DC3000}(t), CP_{t_{r_1}}\left(CP_{t_{b_1}}(K_0, k_{mock}(t,t')), K_0\right)\right).$$

An example branch-recombinant expression profile is shown in Figure 2(b) in the main text. We note that relatively few genes were identified as having their expression levels fully converge back to control levels. Of those that did, none were identified as being 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 several 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 hours post infection (hpi) the impact of effectors was evident by the number of GOs identified, with processes associated with nuclear processes, in particular

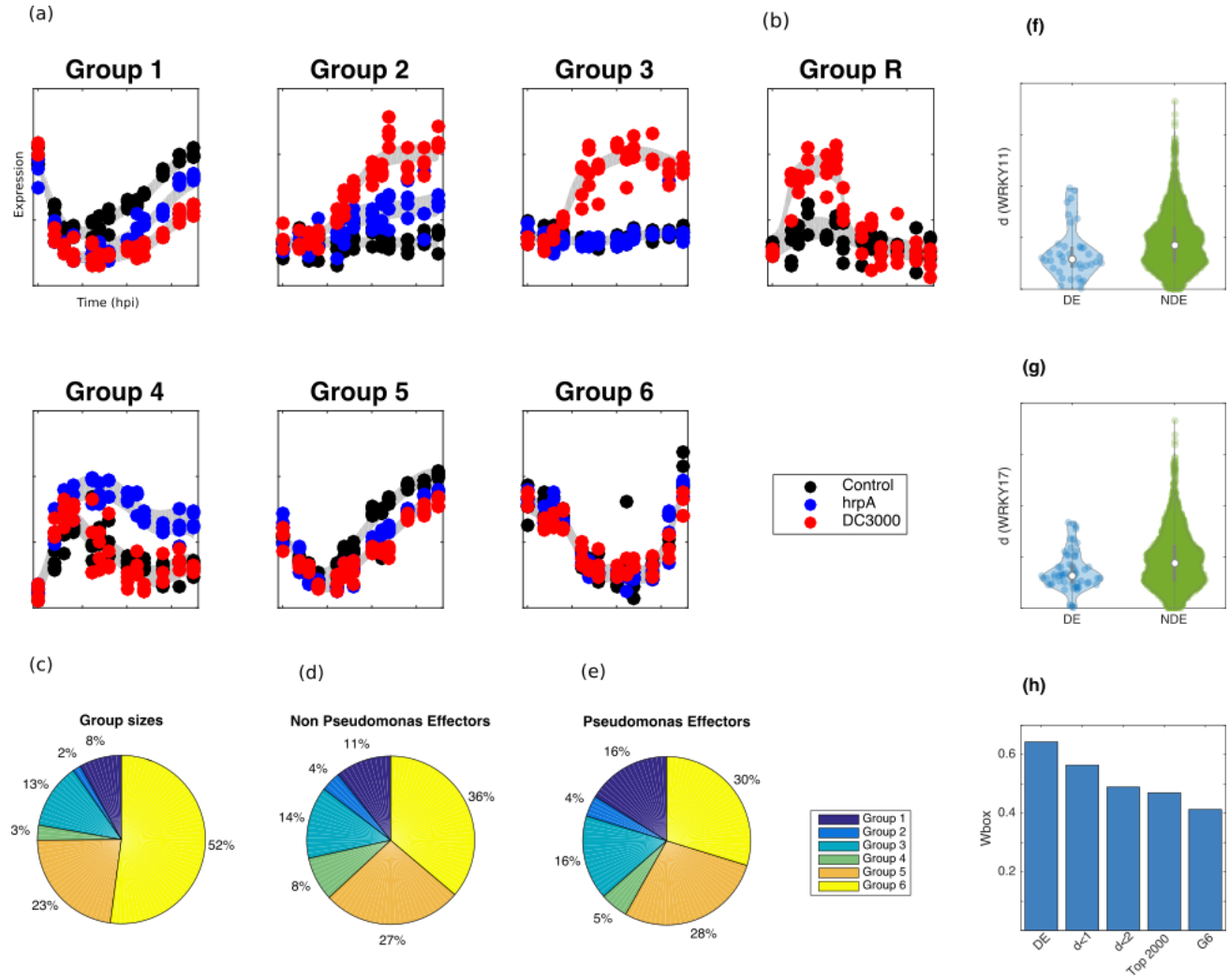


Figure 2: **Branching processes were fitted to the three *Arabidopsis* time series, with hyperparameters optimised to MAP values, and the BIC used to select optimal branching structure. (a)** Example expression profile plots for each of the different classes of branching. **(b)** Example expression profile 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.

chromatin remodelling, nuclear transport, and transcription, most highly enriched. Other GO processes, such as hormone responses and primary metabolism, were, unexpectedly, less abundant. While Lewis et al. (2015) also reported evidence for chromatin remodelling in this dataset, B-RGPs provided 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 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 targets of non-*Pseudomonas* effectors, hypothesising that the Arabidopsis immune response to different pathogens might be conserved (Mukhtar et al. 2011). Figure 2(c, d), shows that these groups were indeed enriched. Next, we checked

for enrichment of Avr and Hop effectors that are present in several strains of *Pseudomonas syringae*, and were again enriched in DC3000-responsive groups (Figure 2(c, e)).

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 7 we indicate the frequency of 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.

Finally, we investigated whether inferred branch times of key regulators were predictive of branching of the direct and indirect targets of key regulators. Here we focused on WRKY11 and WRKY17, known to be amongst the earliest branching transcription factors (TFs) 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 suppression by DC3000 effectors. Genes that branched between control and hrpA and between hrpA and DC3000 were assigned a Euclidean distance (d) based on the position of their branch times with respect to 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 2(f, g)), suggesting that genes that branched at similar times to WRKY11/17 were likely to represent a core set of genes targeted by the pathogen's virulence strategy. 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 (see Supplementary Figure 8); (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 identified as DE in the WRKY11 knockout line compared to control, 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 2(h); see also Supplementary Table 2). Altogether, these results suggest that estimation of branch times may be useful for identifying direct and indirect targets of perturbed genes, and more generally demonstrate the efficacy of B-RGPs for extracting temporally resolved information from complex biological datasets.

5. Discussion

The ability to identify and quantify branching and recombination processes from systems-level measurements has a variety of important applications in the biological sciences. Here we applied B-RGPs to a time-series microarray data of *Arabidopsis thaliana* infected with a bacterial pathogen *Pseudomonas syringae*. B-RGPs provided better temporal resolution of the dynamic host-pathogen interactions, reflecting the suppression of host defences and numerous metabolic changes. Using these approaches, we were able to re-examine the critical initial 8 hours post pathogen inoculation and analysed how the general cellular processes predict a transition from initial enhanced proteolytic activity concomitant with establishment of the capacity to gear up *de novo* protein synthesis, both in terms of ribosome complexes and availability of amino acid precursors. Subsequent enrichment of chromatin remodelling processes predicts a phase of large-scale transcriptional reprogramming. This is consistent with bacterial multiplication phase 5-8 hpi where energy-dependent processes necessary to sustain the pathogen virulence strategy are strongly over-represented. Furthermore, the ability to automatically separate gene expression patterns according to the number branches allowed us to identify subsets of genes enriched for targets of effector proteins, whilst the ability to infer branch times proved insights into the direct and indirect targets of key early transcription factors.

More generally, B-RGPs represents a flexible approach for the analysis of branching and recombination in time series datasets, and can be thought of as a natural extension to two-sample based approaches the analysis of arbitrary numbers of time series.

Whilst here we focused on branching as a function of time, our approach is equally amenable to branching as a function of any other variable, such as expression level of a specific regulator. An intriguing possibility is therefore to incorporate B-RGPs into existing GP-based approaches for the inference of nonlinear dynamical systems (Penfold and Wild 2011; Penfold et al. 2012; Penfold, Millar, and Wild 2015; Penfold et al. 2015; Aijo and Lahdesmaki 2009), which would naturally allow inference of nonstationary nonlinear dynamical systems, such as temporally or spatially varying networks.

In addition, we envisage that B-RGPs could also be useful to capture transcriptional dynamics underpinning cell fate decisions from single cell transcriptomics data. For this, cells are first pseudotemporally ordered along a developmental axis using a combination of dimensionality reduction techniques and curve-fitting or graph-theoretic approaches (Trapnell et al. 2014; Bendall et al. 2014; Marco et al. 2014; Ji and Ji 2016; Setty et al. 2016). Once ordered along pseudotime, B-RGPs could capture the branching dynamics of individual genes, thus identifying the earliest molecular events controlling cell fate decisions. Alternatively, B-RGPs could be used to directly model cell fate decisions. Recent studies by (Reid and Wernisch 2016) have shown how Gaussian process latent variable modes (GPLVMs), can be used to pseudotemporally order genes along a developmental axis, with a key advantage over other pseudotime approaches: the incorporation of capture time into the inference procedure. However, due to a previous lack of treatment for branching in GO models, the approach of Reid and Wernisch (2016) did not explicitly allow for pseudotemporal ordering of datasets with branching behavior. The incorporation of B-RGPs into a GPLVM model would naturally allow for pseudotemporal ordering over branching process, whilst retaining the ability to leverage highly informative data, such as capture time. In our accompanying paper (Sybirna *et al.*, 2018, submitted; attached as supplementary) we demonstrate the use of B-RGPs for analysis of transcriptional branching in early human development.

Acknowledgements

We thank Ufuk Günesdogan and Naoko Irie for critical reading of the manuscript, and the Surani lab for useful insights. We thank Sabine Dietmann for early discussions about branching processes. We also 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 Scholarship and Cambridge International Trust Scholarship. MAS is supported by HFSP and a Wellcome Trust Senior Investigator Award. JR and LW are funded by the UK Medical Research Council (Grant Ref MC_U105260799). YH is supported by a studentship from the James Baird Fund, University of Cambridge. MG acknowledges funding from 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.

Conflict of Interest: none declared.

References

Aijo, T., and H. Lahdesmaki. 2009. 'Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics', *Bioinformatics*, 25: 2937-44.

Bendall, S. C., K. L. Davis, A. D. Amir el, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er. 2014. 'Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development', *Cell*, 157: 714-25.

Boller, Thomas, and Sheng Yang He. 2009. 'Innate Immunity in Plants: An Arms Race Between Pattern Recognition Receptors in Plants and Effectors in Microbial Pathogens', *Science*, 324: 742-44.

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.

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.

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

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

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

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.

Ji, Z., and H. Ji. 2016. 'TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis', *Nucleic Acids Res*, 44: e117.

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

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.

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*.

Marco, E., R. L. Karp, G. Guo, P. Robson, A. H. Hart, L. Trippa, and G. C. Yuan. 2014. 'Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape', *Proc Natl Acad Sci U S A*, 111: E5643-50.

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.

Penfold, C. A., V. Buchanan-Wollaston, K. J. Denby, and D. L. Wild. 2012. 'Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks', *Bioinformatics*, 28: i233-41.

Penfold, C. A., J. B. Millar, and D. L. Wild. 2015. 'Inferring orthologous gene regulatory networks using interspecies data fusion', *Bioinformatics*, 31: i97-105.

Penfold, C. A., A. Shifaz, P. E. Brown, A. Nicholson, and D. L. Wild. 2015. 'CSI: a nonparametric Bayesian approach to network inference from multiple perturbed time series gene expression data', *Stat Appl Genet Mol Biol*, 14: 307-10.

Penfold, C. A., and D. L. Wild. 2011. 'How to infer gene networks from expression profiles, revisited', *Interface Focus*, 1: 857-70.

Poincaré, H. 1885. 'Sur l'équilibre d'une masse fluide animée d'un mouvement de rotation', *Acta mathematica*, 7: 259-380.

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.

Setty, M., M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er. 2016. 'Wishbone identifies bifurcating developmental trajectories from single-cell data', *Nat Biotechnol*, 34: 637-45.

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.

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.

Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. 2014. 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nat Biotechnol*, 32: 381-86.

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.

Zawadzka, M., L. E. Rivers, S. P. Fancy, C. Zhao, R. Tripathi, F. Jamen, K. Young, A. Goncharevich, H. Pohl, M. Rizzi, D. H. Rowitch, N. Kessaris, U. Suter, W. D. Richardson, and R. J. Franklin. 2010. 'CNS-resident glial progenitor/stem cells produce Schwann cells as well as oligodendrocytes during repair of CNS demyelination', *Cell Stem Cell*, 6: 578-90.