

# Profiling Primitive Haematopoietic Progenitors in the Vertebrate Embryo.

---

Lucy S. Wheatley  
Candidate number: 43311  
St. Cross College

**DPhil. Thesis**

**Word count: 40,445**

## Abstract

The primitive wave of haematopoiesis provides an initial circulatory system used during vertebrate embryonic development. During the primitive wave, lateral plate mesoderm cells, expressing the transcription factor *Scl*, contribute to both blood and vasculature progenitors. *Scl* protein is required for the development of all blood lineages in zebrafish, while in mice *Scl* knockout results in prenatal fatalities due to complete absence of early haematopoiesis. The highly dynamic nature and small cell number of primitive haematopoietic progenitors have to date hindered in depth *in vivo* studies of this developmentally crucial population.

Using genome-wide profiling of *scl*-expressing progenitors in zebrafish I have shown that early primitive haematopoietic programmes at the anterior and posterior of the embryos, are distinct at the level of both transcriptional and chromatin landscapes. I have identified characteristic anterior and posterior transcriptional signatures and associated putative *cis*-regulatory modules, correlating with divergent biological functions in these populations.

In particular, I have characterised the cellular heterogeneity, identifying spatially and transcriptionally distinct sub-populations within the anterior *scl*-expressing cells *in vivo*. I have identified regulatory programmes underlying development of anterior vascular and haematopoietic progenitors, and identified a cell subpopulation co-expressing key regulators of both lineages, possibly accounting for the developmental plasticity within the system.

Here, *scl*-expressing haematopoietic progenitor populations were profiled at unparalleled temporal and spatial resolution, obtaining full genome-wide description of early vertebrate primitive haematopoiesis *in vivo*. This work provides comprehensive framework for analysis of gene regulatory interactions and exploration of novel factors and putative regulatory elements involved in this process.

# Index

Abstract.....	ii
Index.....	ii
1. Introduction.....	1-0
1.1. The haematopoietic and vascular system in vertebrates.....	1-0
1.1.1. Embryonic origins of the circulatory system.....	1-0
1.1.2. Primitive and Definitive haematopoiesis.....	1-4
1.1.3. Molecular players involved in haematopoietic and vascular development.....	1-5
1.2. The Scl/Tal1 protein.....	1-7
1.2.1. Discovery of a key leukemic factor.....	1-7
1.2.2. Scl in different biological contexts.....	1-9
1.2.3. Scl binding partners and complex formation.....	1-12
1.2.4. Diverse molecular functions of Scl <i>in vivo</i> .....	1-15
1.2.5. Regulation of Scl activity.....	1-19
1.3. The zebrafish model.....	1-22
1.3.1. Zebrafish general development.....	1-22
1.3.2. Zebrafish model of haemangiogenesis.....	1-24
1.3.3. Developmental origins of the <i>scl<sup>+</sup></i> population.....	1-27
1.3.4. Zebrafish <i>scl</i> isoforms.....	1-30
1.4. Overview.....	1-32
2. Materials and Methods.....	2-35
2.1. Molecular techniques.....	2-35
2.1.1. BAC transgenesis.....	2-35
2.2. Fish protocols and husbandry.....	2-41
2.3. Expression analysis.....	2-42
2.3.1. Transcript detection.....	2-42

2.4.	Sample Preparation.....	2-45
2.5.	Bioinformatic analysis.....	2-48
2.5.1.	Generation of <i>scf</i> specific transcriptomic data.....	2-48
2.5.2.	ATAC-Seq.....	2-49
2.5.3.	Single Cell RNA Sequencing.....	2-50
2.6.	Imaging techniques.....	2-52
2.7.	Primer table.....	2-53
3.	Generation and testing of transgenic fish lines.....	3-0
3.1.	Generation of BAC transgenic fish lines.....	3-0
3.1.1.	BAC selection.....	3-0
3.1.2.	Construct structure.....	3-1
3.1.3.	BAC transgenesis.....	3-3
3.1.4.	Screening for germline transmission.....	3-6
3.1.5.	Confirmation of transgene expression <i>in vivo</i> .....	3-8
3.2.	Characterization of transgenic fish line.....	3-9
3.2.1.	Characterisation of transgene expression <i>in vivo</i> .....	3-9
3.3.	Preliminary biological sample collection and optimisation.....	3-13
3.3.1.	Choice of time points.....	3-13
3.4.	Chapter 3 summary.....	3-16
4.	Investigating early transcriptional variation within <i>in vivo scf<sup>+</sup></i> populations.....	4-18
4.1.	Introduction.....	4-18
4.1.1.	Overall expression levels of <i>scf<sup>+</sup></i> contexts.....	4-19
4.1.2.	Uses and limitations of gene ontology analysis.....	4-22
4.2.	Spatial variation in expression at the 10ss.....	4-23
4.2.1.	Validation of RNA-seq datasets.....	4-24
4.2.2.	Anterior enriched genes.....	4-26
	Posterior enriched genes.....	4-28
4.2.3.	.....	4-28

4.2.4.	Common expressed genes in the 10ss spatial <i>sc<sup>l+</sup></i> comparison .....	4-31
4.3.	Temporal transcriptome variation in the posterior- 10ss vs. 20ss .....	4-33
4.3.1.	10ss enrichment in <i>sc<sup>l+</sup></i> posterior comparison .....	4-33
4.3.2.	20ss enrichment in <i>sc<sup>l+</sup></i> posterior comparison .....	4-35
4.3.3.	Common factors from the <i>sc<sup>l+</sup></i> posterior comparison.....	4-38
4.4.	Common <i>sc<sup>l+</sup></i> cell factors.....	4-40
4.5.	Enriched <i>sc<sup>l+</sup></i> expression profiles compared to <i>sc<sup>l</sup></i> neighbouring cells.....	4-45
4.5.1.	Differential expression between <i>sc<sup>l+</sup></i> and neighbouring <i>sc<sup>l</sup></i> contexts .....	4-46
4.5.2.	10ss Anterior.....	4-48
4.5.3.	10ss Posterior .....	4-60
4.5.4.	20ss Anterior.....	4-70
4.5.5.	20ss Posterior .....	4-83
4.6.	Chapter 4 summary.....	4-0
5.	Investigating early variation in chromatin accessibility within <i>in vivo sc<sup>l+</sup></i> populations.....	5-0
5.1.	Introduction.....	5-0
5.2.	Initial differences in chromatin accessibility in early <i>sc<sup>l+</sup></i> cell populations. ....	5-1
5.3.	Analysis of regulatory loci associated with context specific-ATAC peaks .....	5-3
5.3.1.	Temporal comparison of genes associated with non-promoter ATAC peaks .....	5-4
5.3.2.	Spatial comparison of genes associated with non-promoter ATAC-peaks.....	5-6
5.3.3.	Analysis of binding site motif enrichment within with context specific-ATAC peaks	5-7
5.4.	Chapter 5 summary.....	5-11
6.	Investigating early anterior diversity of <i>sc<sup>l</sup></i> expressing cells .....	6-15
6.1.	Introduction.....	6-15
6.2.	Visualising <i>sc<sup>l</sup></i> sub-populations <i>in vivo</i> .....	6-16
6.2.1.	<i>sc<sup>l</sup></i> populations related to the lateral plate mesoderm.....	6-17
6.2.2.	<i>sc<sup>l+</sup></i> populations in relation to endothelial progenitor cells.....	6-23
6.3.	Transcriptional profiles of individual anterior <i>sc<sup>l+</sup></i> cells at the 10ss. ....	6-29

6.3.1. Overview of single cell sequencing results.....	6-29
6.3.2. Early anterior subpopulation transcriptomes.....	6-31
6.3.3. Transcriptomic diversity of 10ss anterior <i>sc<sup>+</sup></i> cells based on bulk RNA-seq.....	6-37
6.4. Chapter 7 summary.....	6-40
7. Discussion.....	7-45
8. Glossary.....	8-59
9. List of Figures.....	9-60
10. List of Tables.....	10-63
11. Bibliography.....	65

# 1. Introduction

## 1.1. The haematopoietic and vascular system in vertebrates

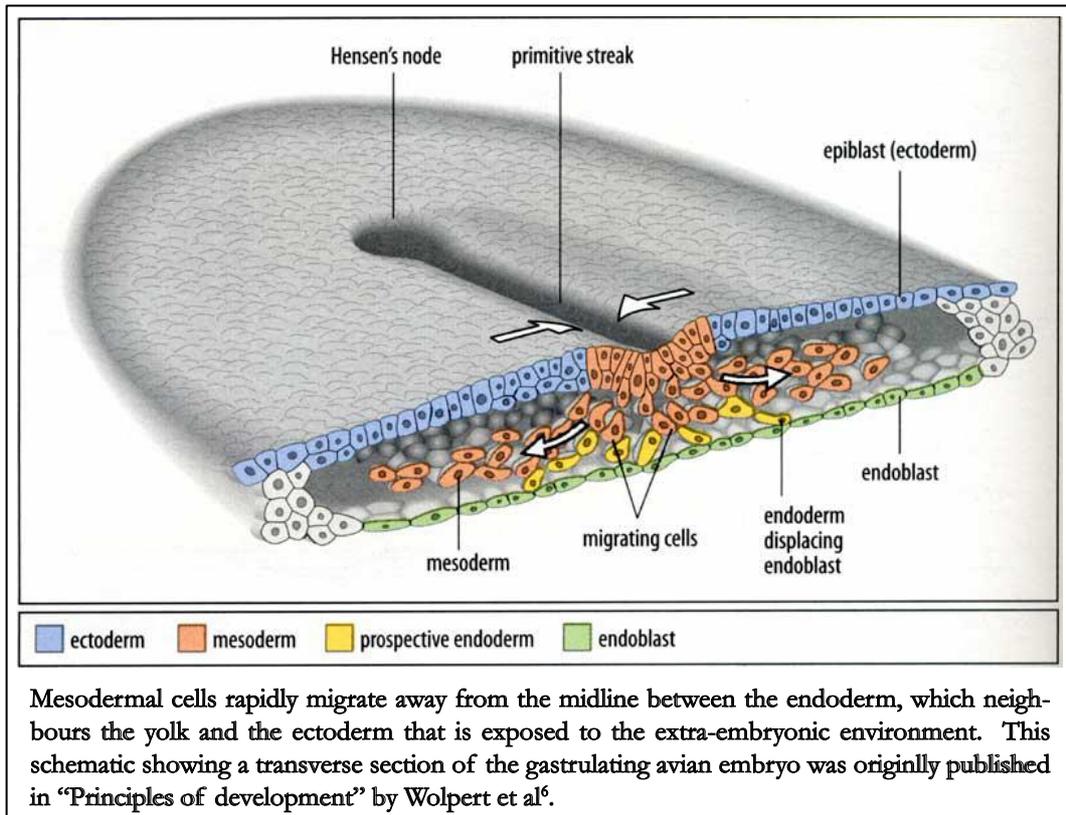
### 1.1.1. Embryonic origins of the circulatory system.

The circulatory system is the first system to arise and function in the developing vertebrate embryo, with high conservation across vertebrate taxa. Circulation permits efficient exchange of nutrients and waste products while also providing a rapid communication system between distant tissues as well as immune defences.

Historically, the avian models were used in the investigation of early vertebrate development. Chick embryos are relatively large, readily available and could easily be manipulated during early development of key tissues such as the blood and vasculature. Thus many of the key findings concerning vascular and haematopoietic development were originally made in the chick or quail.

In gastrulation, the three germ layers: ectoderm, mesoderm and endoderm are first formed in the early embryo, as shown in figure 1-1. During the earliest stages of mesoderm formation, this germ layer consists of highly migratory fibroblast-like cells<sup>1,2</sup>. The extra-embryonic yolk sac is formed from mesodermal cells that rapidly migrate out from the midline, between the ectoderm and endoderm cell layers<sup>3,4</sup>. The extra-embryonic coelom advances and laterally splits the mesodermal population, forming the splanchnopleuric and somatopleuric mesoderm. The splanchnopleuric mesoderm is the ventral layer bordered by the coelom above and the endoderm below, while the somatopleuric mesoderm lies above the coelom and beneath the ectoderm. The first vascular structures that appear during vertebrate development are the blood islands<sup>5</sup>. Proximity to the endoderm and ectoderm influence the cellular fate of the mesodermal cells that migrate between them.

Figure 1-1- Schematic of mesoderm formation, by Wolpert and Tickle, 2010<sup>6</sup>.



These clusters of haemangiogenic cells (cells capable of forming haematopoietic and vascular cells) develop from the splanchnopleuric mesoderm suggesting repressive signalling arises from the ectoderm while positive vasculogenic signal emanate from the endoderm. As a result of these gradients of positive and negative intercellular signals the haemangiogenic clusters are correctly spatially orientated and restricted. Haematopoietic precursors were originally proposed to arise from the centre of the blood island, while the mesodermal cells on the periphery of the blood islands form the angioblast population (endothelial precursors)<sup>7</sup>. Hence the close spatial proximity of these two cell types led to the proposition of the existence of a common blood and vascular progenitor<sup>7</sup>. The cells at this site of the potential origins of the blood and vascular lineages was called the haemangioblast, though at this stage Murray was unable to conclude whether this referred to a mixed population or a bi-potent progenitor population<sup>7,8</sup>. Grafting studies in the mouse embryo suggested that vascular and haematopoietic lineages arise independently<sup>9</sup>. Clonal analysis of

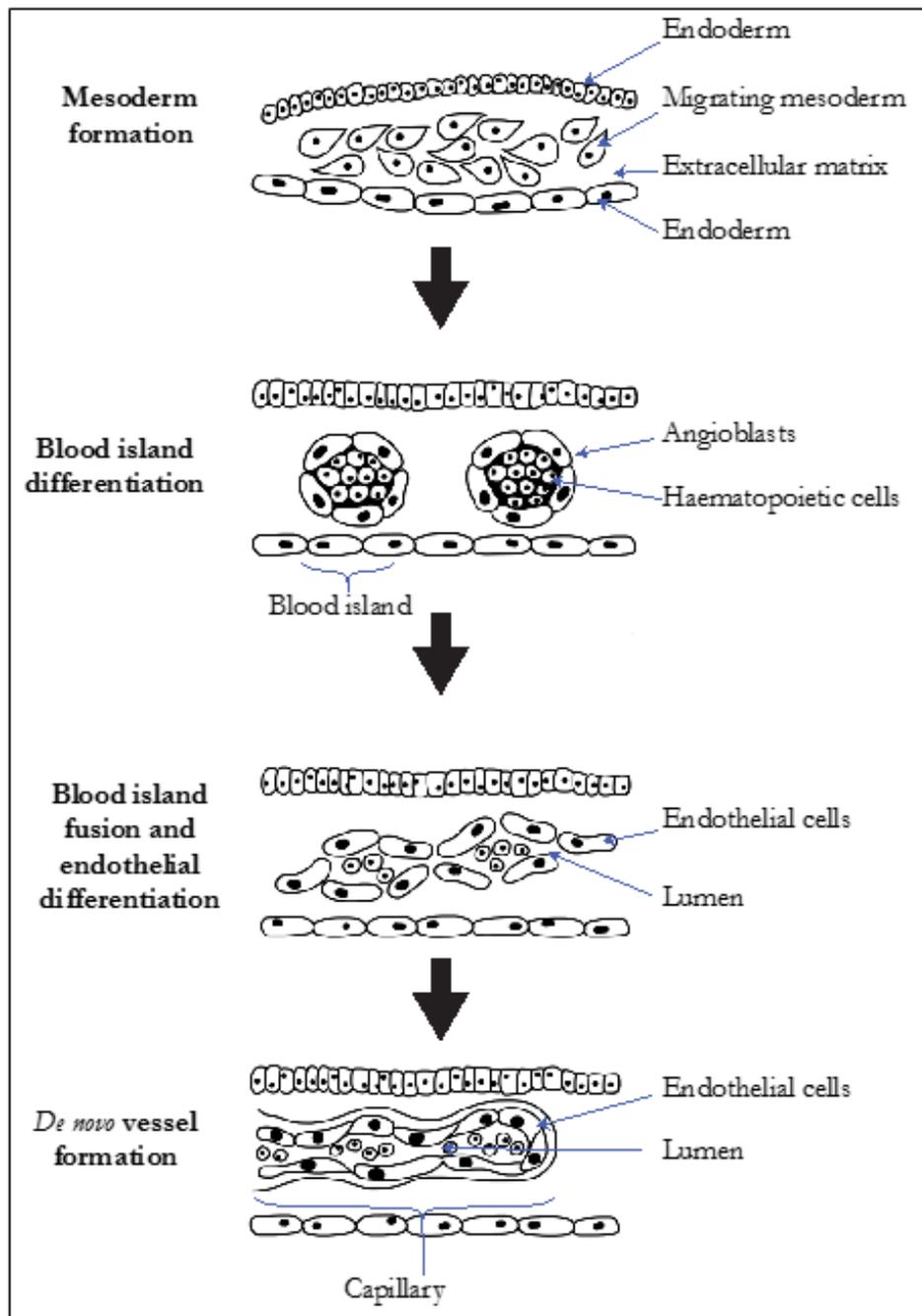
differentiated mouse embryonic stem cells identified a blast colony forming cell (BL-CFC), a single cell that could be differentiated *in vitro* to give rise to either haematopoietic or endothelial lineages<sup>10</sup>. However this study represented a forced differentiation rather than observation of a naturally occurring event, thus may not even occur *in vivo* and still did not provide any information on the embryonic origin of haematopoiesis. Smooth muscle cells were also observed to differentiate from BL-CFCs suggesting that these cells represent an earlier mesodermal progenitor rather than a more restricted bi-potential haemangioblast. In the primitive streak of the mouse embryo, rare BL-CFC like cells were identified, however their isolation and subsequent culture only gave rise to haematopoietic lineages<sup>11,12</sup>.

Some studies suggest that committed haemangiogenic precursors exist prior to gastrulation, but fail to clarify whether both haematopoietic and vascular functions arise from a single cell *in vivo*, or whether two distinct populations exist spatially intermingled<sup>13-15</sup>. In an attempt to resolve this transgenic mouse embryos were subjected to random single cell labelling prior to gastrulation, the resultant clones observed were haematopoietic, endothelial or rarely a mix of both lineages<sup>16</sup>. The authors discussed these results as counteracting the idea of the haemangioblast, with mixed clones originating from a pre-gastrulation precursor or a late stage endothelial to haematopoietic transition.

Supporting the existence of a haemangioblast in vertebrates is the shared expression of key molecular markers in both developing blood and endothelial cells. Overexpression of certain factors that are common to both fates has show an increase in both haematopoietic and vascular numbers, however this is often accompanied by expansion of other fates and does not mimic endogenous development<sup>17-20</sup>. This has been further supported by fluorescent labelling of single

cells and cell tracing studies within the zebrafish, which identified cells capable of contributing to both haematopoietic and vascular, but not other mesodermal fates<sup>21</sup>.

Figure 1-2- Schematic of vertebrate vasculogenesis, modified from Risau and Flamme, 1995<sup>22</sup>.



The embryonic origins of the haematopoietic system and whether this involves transition through a haemangioblast still remains unclear<sup>23</sup>. Knowledge of these key developmental processes is essential for informing cellular reprogramming and regenerative clinical approaches. The main obstacle to tackling this question has

been our inability to study these progenitors *in vivo* given their relatively small numbers within the embryos, combined with the failure to fully recapitulate their development in culture.

As the somites begin to form, the first angioblasts are observed within the embryo proper, in close proximity to the endoderm and begin to fashion the earliest version of the dorsal aorta<sup>24</sup> (as detailed in figure 1-2). As with the haematopoietic lineage, the origins of angioblasts in the embryo proper have also been controversial. Due to the inherently dynamic and motile nature of mesodermal cells, tracing the origins of mesodermal populations can be challenging, at the earliest stages of development. The intra- and extra-embryonic vessels connect shortly after the formation of the second somite and cephalic mesodermal angioblasts stream into the head where they begin formation of the cranial vasculature.

Two distinct yet interconnected biological processes form blood vessels. Initially vasculogenesis occurs by aggregation of angioblasts, resulting in the *de novo* formation of a vessel at the location of the aggregate<sup>25</sup>. Angiogenesis occurs subsequently to vasculogenesis and this process consists of sprouting or splitting off existing vessels in response to signals from neighbouring tissues<sup>26</sup>.

### 1.1.2. Primitive and Definitive haematopoiesis

In all vertebrates, haematopoietic development occurs in two key waves: primitive and definitive. Primitive haematopoiesis provides an initial circulatory system for embryonic development. This is superseded by the definitive wave of haematopoiesis, which produces self renewing haematopoietic stem cells (HSC) capable of developing into all mature haematopoietic cell types<sup>27,28</sup>. In vertebrates primitive and definitive haematopoiesis takes place in different embryonic locations. Although these locations vary between different vertebrate species, molecular

signatures and many of the regulatory pathways involved in haematopoietic development are highly conserved across the vertebrate taxa.

### 1.1.3. Molecular players involved in haematopoietic and vascular development

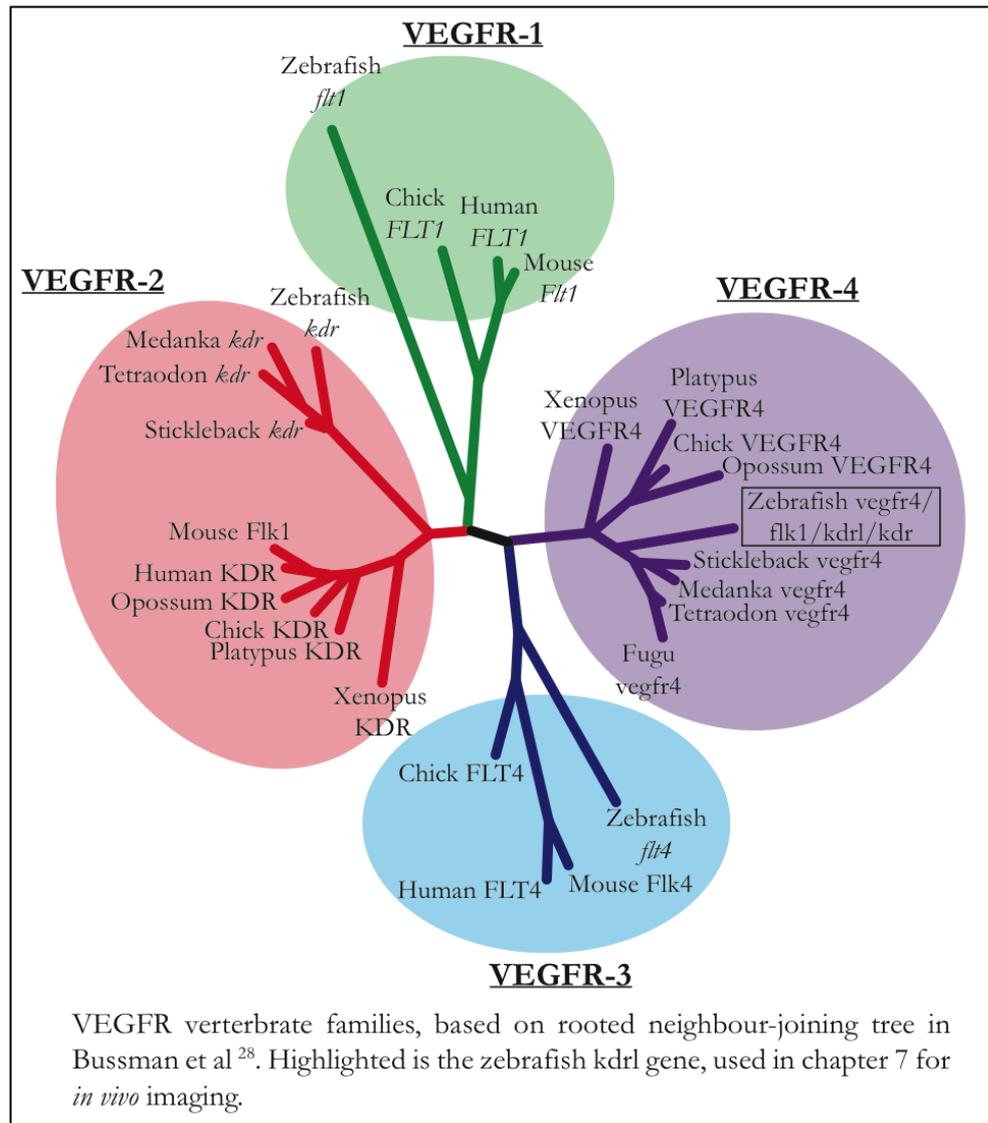
A key signalling pathway regulating the formation of vasculature is mediated by Vascular Endothelial Growth Factor (VEGF) receptors and ligands. Overexpression of VEGF ligand alone in the avian embryo leads to ectopic vasculogenesis<sup>29</sup>. VEGFR2 (VEGF receptor 2) was identified as one of the first molecular markers expressed in angioblasts in both mice and quail, with its expression becoming restricted to endothelial cells as development progresses<sup>29-32</sup>. Figure 1-3 details the evolution of the VEGF-R family in vertebrates used by Bussman et al to clarify the roles of these proteins in zebrafish<sup>33</sup>.

Complete disruption of *vegf* or *vegfr* (VEGF receptors) genes in mice is lethal early in development due to the crucial role of these factors in vasculogenesis and later in angiogenesis<sup>34,35</sup>. VEGF signalling has been shown to act upstream of Notch signalling and downstream of Sonic hedgehog activity in arterial lineage signalling<sup>36</sup>. VEGF signalling also increases vascular permeability<sup>37</sup>. Engerman *et al.* showed that under normal conditions endothelial cells rarely undergo proliferation in adult tissue, however during development endothelial cells rapidly proliferate in correlation with the expression of VEGF and VEGFRs<sup>38,39</sup>.

ETS family of proteins play crucial yet overlapping roles in blood vessel development and myeloid lineage formation<sup>20,40-43</sup>. This transcription factor family bind to the sequence (C/A)GGA(A/T) (reviewed by Sharrocks *et al.*<sup>43</sup>) and includes *etv2*, *erg* and *fli1* genes. Knockdown of *ets* family genes results in defects in myelopoiesis, vasculogenesis and angiogenesis, suggesting these molecules are involved in all three developmental processes<sup>20,44,45</sup>. For instance, *etv2* knockdown

causes dramatic reduction in cell numbers of macrophages, neutrophils and heterophilic granulocytes<sup>41</sup> as development progresses in those mutants. ETV2 protein plays a critical role in the formation of vasculature as knockdown also causes complete loss of circulation<sup>42</sup>. Moreover, *etv2* and *foxc1a* double knockdown results in loss of the majority of developing vasculature<sup>46</sup>.

Figure 1-3- Vertebrate VEGFR classes, adapted from Bussmann *et al.* 2007<sup>47</sup>.



Some of the earliest markers that have been used to probe angioblast development are cell adhesion factors, such as VE-cadherin (CDH5), platelet endothelial cell adhesion molecule (PECAM-1) and cluster of differentiation in a cell surface glycoprotein *cd34*<sup>48-51</sup>. *Cdh5* has been shown to play a crucial role in mediating VEGF signalling required for sprouting angiogenesis<sup>52,53</sup> and for controlling blood

vessel permeability<sup>54</sup>. Morpholino knockdown in zebrafish identified an angiogenic-independent role for *cdh5* in cardiac development: its depletion resulted in low cardiac output due to poor endocardial cell adhesion<sup>55</sup>.

Rbtn2, a factor identified through its contribution to T-cell Acute Lymphoblastic Leukaemia T-ALL, also known as Lmo2<sup>56</sup>, plays crucial roles in vascular and haematopoietic development. Lmo2 is a LIM-only domain protein that cannot directly bind DNA<sup>56-58</sup> and requires association with other proteins, such as Stem cell like protein, Scl (also known as T-cell Acute Lymphocytic Leukaemia 1, Tal1), to stabilize its association to DNA<sup>59</sup>. Similar to *scl*, *lmo2* is expressed in developing blood and vasculature cells of vertebral embryos.<sup>60-63</sup>

## 1.2. The Scl/Tal1 protein.

### 1.2.1. Discovery of a key leukemic factor.

SCL (Stem Cell Leukemia) was first discovered as a gene involved in a chromosomal translocation associated with the occurrence of the stem cell-like acute lymphoblastic leukemia (ALL), presenting both a myeloid and lymphoid differentiation phenotype<sup>63-65</sup>. Leukemic cells from a 16-year old patient showed an early T-cell phenotype, defined as being CD2<sup>+</sup>/CD7<sup>+</sup> and CD3<sup>-</sup>/CD4<sup>-</sup>/CD8<sup>-</sup>. When typical chemotherapeutic treatments failed to be effective in this patient, adenosine deaminase inhibitors, such as 2'-deoxycoformycin, were administered, as they were previously shown to treat leukemic growth by blocking nucleotide synthesis. During the first 7 days of 2'-dexycoformycin treatment the patient's leukemic cells dramatically changed morphology from early T-lymphoid-like to displaying myeloid morphological features. At the same time, the expression of CD7, marker of early T-lymphocytes, was lost, resulting in leukemic cells displaying classical stem-like features<sup>66</sup>.

Kurtzberg *et al.* showed that a similar “stem cell like” phenotype could be generated *in vitro* with 2'-deoxycoformycin treatment of primary patient leukemic cells, resulting in cells differentiating down a range of lineages<sup>65</sup>. During this conversion, the genomes of the leukemic cells were monitored and showed a chromosomal translocation occurring between chromosome 1p32-33 and the *TCR*  $\alpha/\delta$  locus on chromosome 14q11.

The *SCL* gene has hence been identified and mapped to human chromosome 1 by analysing this translocation and through independent efforts of several groups<sup>64,67</sup>. Aberrant recombinase activity was suggested to be responsible for this chromosomal rearrangement, as sequences found in proximity to the human *SCL* gene show high similarity to specific DNA sequences flanking the *TCR*  $\alpha/\delta$  locus<sup>64,67</sup>. It has thus been suggested that these motifs are used by recombination enzymes to correctly identify and arrange segments of DNA for functional *TCR* gene assembly. These putative recombinase signal sequences occur multiple times through the human *SCL* locus, but are not conserved in mice<sup>68-70</sup>. The rearrangement of the *TCR* locus occurs at a stage of development when *SCL* is highly expressed in haematopoietic cells; thus the *SCL* locus is likely to be contained in hypomethylated chromatin<sup>71</sup>. This “open” chromatin state results in exposed DNA and therefore may be permissive to erroneous recombinase activity at this site. In this initial T-ALL case, the intact *SCL* coding region was relocated into the *TCR* locus, resulting in inappropriate expression of the *SCL* protein under the control of *TCR* regulatory elements<sup>64</sup>. Consequently high *SCL* expression is observed in leukemic T-cells and undetectable in normal T-cells<sup>64,72-74</sup>.

A second mechanism of *SCL* erroneous regulation was identified in leukemic cells of other T-cell leukemia patients<sup>75</sup>. Deletion of a ~90kb region of chromosome 1, placed the *SCL* coding region under the regulatory elements of a ubiquitously

expressed gene named *SIL* (*SCL* Interrupting Locus)<sup>76,77</sup>. Further analysis showed that this deletion event frequently occurs in T-cell leukemia patients and is strongly correlated with deletion of at least one of the *TCR δ* genes. This strong correlation is indicative of a role for *SCL* in the process of  $\delta$  locus deletion or commitment to the *TCR α/β* fate<sup>71,78</sup>.

### 1.2.2. *Scl* in different biological contexts

Following the identification of *SCL*'s role in T-ALL, experiments were carried out to determine the biological function and mechanism of action for this gene under normal, non-oncogenic conditions.

Combined with its key role in leukemia, the expression of *Scl* in “early” haematopoietic tissues suggested that *SCL* may act as a key regulator of early haematopoietic fate decisions<sup>64,72,79</sup>. Moreover, levels of *Scl* mRNA were observed to significantly increase when murine erythroleukemia (MEL) cells were induced to differentiate *in vitro*<sup>80</sup>, suggesting a second role in erythroid development for this potential transcription factor.

GATA proteins are transcription factors that bind the DNA sequence (A/T)GATA(A/G). GATA1 was identified in zebrafish following study of a *bloodless* mutation. In this mutation line HSC, myeloid and lymphoid lineages were unaffected yet circulating erythroid cells were lost<sup>81,82</sup>. In mice GATA1 has been shown to be required in the differentiation of erythroid progenitor cells beyond the proerythroblast stage<sup>83</sup>. The promoter of *SCL* was shown to harbour GATA binding sites, that could be bound by a key erythroid factor, GATA1<sup>84</sup>.

Further supporting an early haematopoietic role, overexpression of *SCL* in K562 cells induced spontaneous erythroid differentiation<sup>68,80</sup>. Expression of a dominant negative form of *SCL* in MEL blocked erythroid maturation<sup>80</sup>. *SCL* has been shown

to act both during erythroid specification and maturation, albeit through distinct mechanisms and with different requirements in *SCL* expression levels<sup>85</sup>.

Two groups independently generated *Scf* null mice, and showed they die of severe anaemia by E10<sup>86,87</sup>, with embryos displaying phenotype similar to deficiencies in GATA1<sup>88,89</sup> or Lmo2<sup>63</sup> gene products. A deficiency in myelopoiesis was also described, through investigating the haematopoietic potential of *Scf*<sup>-/-</sup> cells upon injection into mice blastocysts. Despite these cells contributing to a range of cell types, *Scf*-null cells were unable to contribute to haematopoietic lineages, indicating that *Scf* may act at the level of a myelo-erythroid progenitor<sup>86</sup>. Supporting this earlier developmental role, *Gata1* and *Spi1* transcripts were absent from *Scf*<sup>-/-</sup> cells<sup>87</sup>. Further differentiation assays using mutant *Scf*<sup>-/-</sup> cells *in vitro* and in chimeric mice identified a crucial role for *Scf* protein in the development of all haematopoietic lineages including lymphoid lineages<sup>90,91</sup>. During early thymocyte development, *Scf* normally becomes silenced in T-cells<sup>92</sup>.

In addition to expression within haematopoietic tissue, *Scf* was also expressed in the developing vasculature, including in the embryonic yolk sac, however studies in *Scf*-null mice indicated that endothelial development was not ablated<sup>86,87</sup>. To investigate *Scf* functions beyond haematopoietic development, the *Scf* mouse knockout was rescued specifically in haematopoietic cells by expression of *Scf* under the regulatory region of *gata1*<sup>93</sup>. In “rescued” embryos and analysis of chimeras generated from *Scf*<sup>-/-</sup> embryonic stem cells, endothelial cells were shown to be correctly specified, but angiogenic remodelling was severely disrupted and *Scf*<sup>-/-</sup> cells could not contribute to major vessels<sup>93</sup>. However, upon rescue under the control of a *Scf* 3' enhancer that directed expression of *Scf* open reading frame (ORF) to HSCs and endothelium<sup>94</sup>, both haematopoietic and angiogenic defects were resolved<sup>95</sup>. Studies in avian systems confirm that *scf*<sup>+</sup> cells, which arise from the splanchnopleuric mesoderm,

acquire endothelial markers and function in angiogenesis<sup>96</sup>. Morpholino mediated knockdown in zebrafish also determined that *scf* played a crucial role in the formation of the dorsal aorta and major trunk vessels<sup>97</sup>. Conserved regulatory functions for *scf* in both haematopoietic and vascular fates indicated that *scf* might act in the controversial common precursor- the haemangioblast. The existence of a common progenitor is also supported by the evidence from the *cloche* zebrafish mutant line, which displays deficit in both haematopoietic and vascular lineages and is characterised by a complete loss of the endocardium<sup>98</sup>. In *cloche* mutants, *scf* expression is almost completely lost, however ectopic expression of *scf* can partially rescue the haematopoietic and vascular defects, indicating that *scf* lies downstream of *cloche*<sup>99</sup>.

Time-lapse microscopy determined that in the zebrafish embryo, the cells of the developing endocardium express *scf* and originate from the anterior lateral mesoderm (ALM)<sup>47,100</sup>. These cells are initially part of anterior cell population proposed to give rise to early myeloid progenitors, but then rapidly migrate into the developing heart field. In *scf* mutants<sup>101</sup>, where bHLH domain of the Scl protein has been disrupted, these endocardial precursors form but fail to migrate into the heart field, leading to major cardiac defects<sup>47</sup>.

*Scl* is also required to repress the cardiomyogenic program in mice. Upon *Scl* knockdown, endothelial cells of the yolk sac have been shown to ectopically express cardiac markers and spontaneously form beating cardiomyocytes following *in vitro* culture<sup>102</sup>.

In several key vertebrate models, *scf* has been shown to be expressed within neural tissue<sup>72,73,103-106</sup>, displaying highly conserved patterns encompassing thalamus midbrain and hindbrain<sup>104,107</sup>. Hindbrain/spinal cord expression is driven by regulatory modules distinct to those controlling midbrain *scf* expression, in mouse, zebrafish

and chicken models<sup>104</sup>. The regulatory region that controlled *SCL* expression in the midbrain contained two binding sites for GATA transcription factors<sup>104</sup>. Mutation of these binding sites in erythroid cell lines has previously indicated that GATA1 binding was required for full activity of the *SCL* promoter<sup>108,109</sup>. Although *GATA1* is not expressed in the midbrain, *GATA2* and *GATA3* are and their patterns overlap with *SCL* in neural cells<sup>110,111</sup>. Disruption of these GATA binding sites completely ablated neural *SCL* expression as well<sup>104</sup>, indicating that despite different cellular contexts similar regulatory kernels may function and be re-used to direct distinct developmental processes.

Bone homeostasis is the balance of bone formation and reabsorption mediated by osteoblasts and osteoclasts, respectively. Osteoclast development from HSCs has been shown to be under *SCL* regulation and similar factors to those controlling myelopoiesis are also involved<sup>112</sup>. *In vitro*, *SCL* null embryonic stem cells fail to develop into osteoclasts<sup>113</sup>.

It is possible that similar regulatory interactions surround *scf* in each of these biological contexts, with tissue-specific factors tuning *Scf*-mediated activity towards tissue-specific targets.

### 1.2.3. *Scf* binding partners and complex formation.

When the *SCL* gene was identified, the resulting predicted protein showed high homology to previously described basic helix-loop-helix proteins, such as Lyl-1, Myc and MyoD<sup>64</sup>. Murre *et al.* showed that in addition to permitting direct DNA-binding of these related proteins, the helix-loop-helix domain may also contribute to the formation of protein complexes<sup>114</sup>. The high homology of *SCL* protein with factors known to be involved in key developmental fate decisions, suggested that *SCL* also played a crucial developmental role and prompted further investigation. Despite

84% conservation in sequence at the bHLH domain and overlapping expression *in vivo*, Lyl-1 was unable to rescue *Scf* knockout<sup>115</sup>.

As with other bHLH family members, *Scf* was shown to form heterodimers with members of another family of bHLH proteins, E-proteins, including E2A, E47 and E12<sup>116-118</sup>. E-proteins, which are ubiquitously expressed, associate with *Scf* and extend its downstream target pool by permitting it, when in such complexes, to bind to E-box DNA motifs (CAGATG)<sup>117,118</sup>. Direct DNA interactions, in this case, are formed by the basic domain on both *Scf* and its associated E-protein<sup>119</sup>. The crystal structure of *Scf*:E47 interaction domains indicates that the formation of such heterodimers would be thermodynamically more favourable than the formation of *Scf* or E47 homodimers<sup>120</sup>.

*Scf* and *Lmo2* directly interact *in vivo*, and were found associated in T-ALL cell lines and erythroid cells<sup>56</sup>. *Lmo2* knockout mice die at E10.5 due to severe anaemia arising from a block in erythroid maturation<sup>63</sup>, indicating that, along with *Scf*, this protein is essential for proper erythroid development. 80% of childhood T-ALL patients with ectopic expression of either *Scf* or *Lmo2*, show co-overexpression of both genes<sup>121,122</sup>, suggesting a possible co-operative role. These two transcription factors act synergistically to strongly promote tumorigenesis in transgenic mice, causing drastic and early onset of T-cell leukemia that is significantly more severe than in *Lmo2*-only controls<sup>123</sup>. *Lmo2* has been shown to contribute to angiogenesis, but not vasculogenesis. Ectopic *scf* expression in zebrafish embryos resulted in expansion of *lmo2* expression throughout the mesoderm<sup>18</sup>. This expansion of endogenous *lmo2* expression as measured by *in situ* analysis, is greatly increased upon co-injection of *scf* and *lmo2* mRNAs, and is also accompanied by increased expression of haemangiogenic factors along the full length of the anterior/ posterior axis. Despite overexpression of *lmo2* and *scf*, erythropoiesis was only expanded to the pronephric

mesoderm, in a fashion that overlapped with *gata1* expression<sup>18</sup>. Structural studies of the Scl protein:DNA complex revealed that upon Lmo2 binding to Scl:E47 heterodimers, the protein complex is stabilized, but hydrogen bonds with DNA are lost, thus reducing its association with DNA binding motifs<sup>120</sup>.

In *Xenopus*, *Scl* overexpression experiments resulted in expansion of the cellular field of primitive haematopoiesis, with a significantly greater effect achieved upon combined overexpression of *Scl*, *Lmo2* and *GATA1* mRNAs, than with overexpression of any of these factors alone<sup>124</sup>. Scl, the E2A protein E47, and Lmo2 form a multi-protein-DNA-binding complex in erythroid cells, along with GATA1 and Ldb1<sup>125</sup>. This complex was termed the “Scl pentameric complex” and shown to directly associate with E-box-GATA consensus motifs - a DNA sequence commonly found in the promoters of erythroid genes<sup>125,126</sup>. In early mesodermal development GATA1 is absent, thus GATA2, which is expressed at this time, has been proposed to interact with Scl containing complexes and has been shown to directly interact with Lmo2<sup>127</sup>. This is supported by the partial rescue of haematopoietic defects in *GATA1* null mice by ectopic expression of *GATA2*<sup>128</sup>.

Lécuyer *et al.* have shown that Scl multi-protein complexes bind and enhance expression of the *c-kit* promoter *in vivo*<sup>129</sup>. Synergistic action of Scl with other transcription factors suggested that Scl commonly functioned as part of multi-protein complexes with the complex members directing the complex to specific subsets of Scl target genes<sup>130,131</sup>.

Point mutation analysis identified key residues within the bHLH domain essential for haematopoietic function of Scl protein, including residues required for heterodimerization and specifically for Lmo2 association<sup>132</sup>. Mutation of these key residues has emphasised the requirement for Scl to form functional complexes in order to mediate transcriptional regulation<sup>132</sup>.

Surprisingly, the direct DNA binding activity of Scl was found to be dispensable for Scl function *in vitro* and *in vivo*<sup>133,134</sup>. Following disruption of the bHLH region of Scl DNA binding ability was ablated, yet bHLH domain-lacking variants were able to rescue erythroid specification in *scf*<sup>-/-</sup> ES cells. Similarly such DNA binding domain lacking *Scf* variant mRNAs also rescued erythroid and vascular fates in zebrafish *cloche* mutant<sup>85,116,133</sup>, which normally lacks *scf* function. Together, these studies suggest that other proteins complexed with Scl can mediate direct DNA interactions.

Isolation and characterisation of Scl-containing protein complexes in erythroid cultures also identified co-repressors ETO2 and Gfi1b as Scl interaction partners<sup>135</sup>. Thus complex formation could not only determine Scl targets but also the nature of regulatory function exerted upon a bound target.

#### 1.2.4. Diverse molecular functions of Scl *in vivo*.

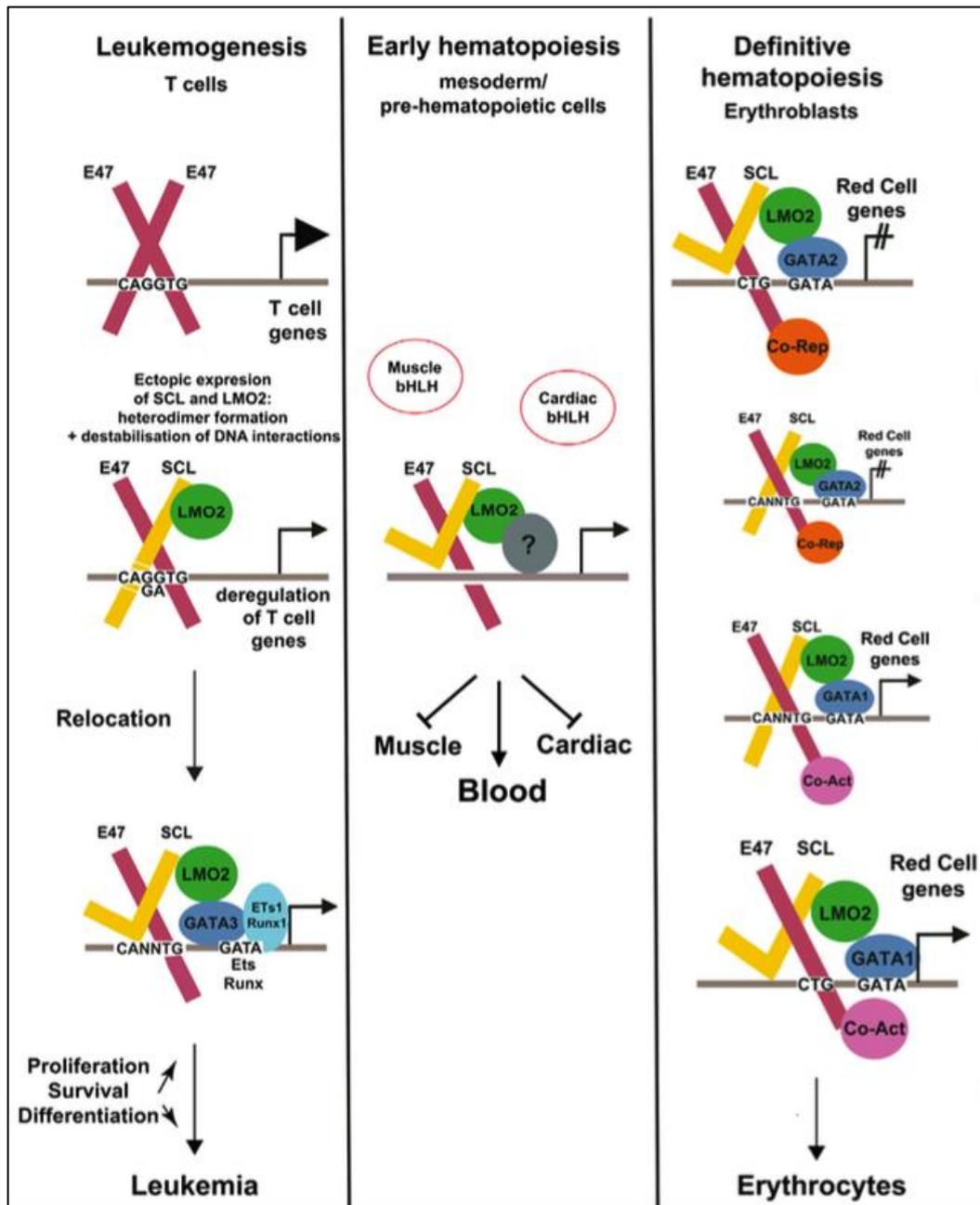
Evidence from different cellular systems indicates that the mechanism of Scl action varies between different cellular contexts. Through the formation of complexes Scl can mediate a range of effects on a variety of target loci. Scl has been shown to recruit a variety of regulatory activities to target gene loci and has been proposed to influence the activity of other transcription factors through modulating the availability of common binding partners.

A range of Scl target loci has been identified through analysis of genome-wide binding profiles of Scl, analysed within a number of different cellular environments by ChIP-Seq (Chromatin Immunoprecipitation followed by deep sequencing)<sup>126,136,137</sup>. For example, in primary erythroid cells, Scl most commonly binds in proximity to GATA motifs to promote expression of downstream targets. Yet in a T-ALL-derived cell-line, Scl binds most frequently near *runx* motifs and mainly results in target gene repression<sup>138</sup>.

Several studies support a sequestration model, theorised to contribute to Scl-mediated leukemogenesis. The model proposed that in T-ALL cells which ectopically express both *Lmo2* and Scl, Scl pentameric complexes continue to form, eventually depleting the free pools of other complex members such as the E-proteins, while at the same time reducing the affinity Scl shows for specific DNA binding sites<sup>120</sup> (detailed in figure 1-4). E-protein homodimers are essential for T-cell maturation, thus in conditions of *Lmo2* and *Scl* overexpression, maturation promoting complexes are removed and replaced with complexes more commonly associated with early haematopoiesis<sup>92,138,139</sup>. Sequestration of E-proteins by Scl-containing complexes has also been reported to contribute to leukemogenesis through preventing apoptotic signalling, which is in part mediated by E-protein homodimers<sup>140</sup>. Scl sequestration of E-proteins has also been proposed to have important indirect effects, as binding with Scl opposes E-protein interaction with bHLH factors involved in the development of other cell types<sup>120</sup>. This theory is supported by *in vivo* studies that describe cardiogenic activity in would-be haematopoietic tissues, following *Scl* knockdown<sup>102,141,142</sup>. Developmental regulation through sequestration of other key factors would thus permit similar regulatory circuitry to be employed in varied biological contexts, i.e. during vascular, blood and neural development. Lineage specific repression would result from the unavailability of E-proteins to associate with tissue-specific bHLH factors such as MyoD and NeuroD.

O'Neil *et al.* used transgenic mice to investigate the effect of Scl interactions with the E-proteins E47 and HEB in leukemogenesis. Upon induced *Scl* expression, mice with reduced E47 or HEB levels showed dramatic disease acceleration compared to wildtypes. E-protein target genes were repressed with co-repressor mSin3A occupying enhancers that in wild type are normally bound by E-protein/p300 complexes and promote expression<sup>143</sup>.

Figure 1-4- Model of Scl protein complex interactions with DNA in different cellular contexts, by El Omari *et al.* 2013<sup>120</sup>.



Scl has been observed to mediate both transcriptional activation and repression upon binding to regulatory regions of a target gene. In the case of *c-kit* Scl binding results in opposing activities and is dependant on the cellular context. Transcription of *c-kit* was promoted by Scl activity in haematopoietic progenitor lines, yet repressed by binding of Scl-containing complexes in primary erythroid cells<sup>144,145</sup>. ETO2-Gfi1b-Scl complexes identified in erythroid cells repress target gene expression *in vitro*.

However, as erythroid maturation progresses, these complexes dissociate, relieving repression of terminal erythroid maturation genes while simultaneously releasing Scl from repressive protein complexes, allowing it to associate to different partners and form activating complexes<sup>135</sup>.

SCL complexes can associate with the broadly expressed transcriptional co-activator p300 in human MEL cells<sup>146</sup>. p300 and CREB-binding protein (CBP) are closely related proteins that associate with transcription factors to increase the transcription of their downstream targets. CBP and p300 act by recruiting basic transcriptional machinery<sup>147-149</sup>, relaxing chromatin structure at the target loci through their intrinsic histone acetyltransferase activity or by recruitment of histone modifiers<sup>150</sup> to transcription factor complexes. This p300:SCL complex was shown to associate with E-box motifs and promote expression of proximal genes during stimulated erythroid differentiation<sup>146</sup>. Conversely in the same cell line SCL protein associates with transcriptional co-repressor, mSin3A, which leads to the recruitment of HDAC1 to SCL targets<sup>151</sup>. HDAC1 catalyses the removal of acetyl groups from histone tails, which promotes chromatin condensation and diminishes transcription locally<sup>152</sup>. In MEL cells this SCL-mediated recruitment of HDAC activity caused transcriptional repression, which was reduced during stimulated erythroid maturation<sup>151</sup>.

SCL can directly interact with LSD1, a histone-3 lysine-4 demethylase that catalyses promoter and enhancer hypomethylation and represses transcription of downstream targets. The direct nature of this interaction indicates that SCL can direct regulatory activity to target genes in the absence of an “SCL core complex”<sup>153</sup>. SCL-mediated recruitment of LSD1 blocks erythroid differentiation through repressing expression of SCL target genes required for terminal differentiation<sup>154</sup>.

The scale of this opposing SCL molecular activity was demonstrated in T-ALL cell culture. Upon knockdown of *SCL*, an array of direct SCL targets were observed to show either an increase or decrease in expression, which culminated in the loss of leukemogenicity<sup>136</sup>.

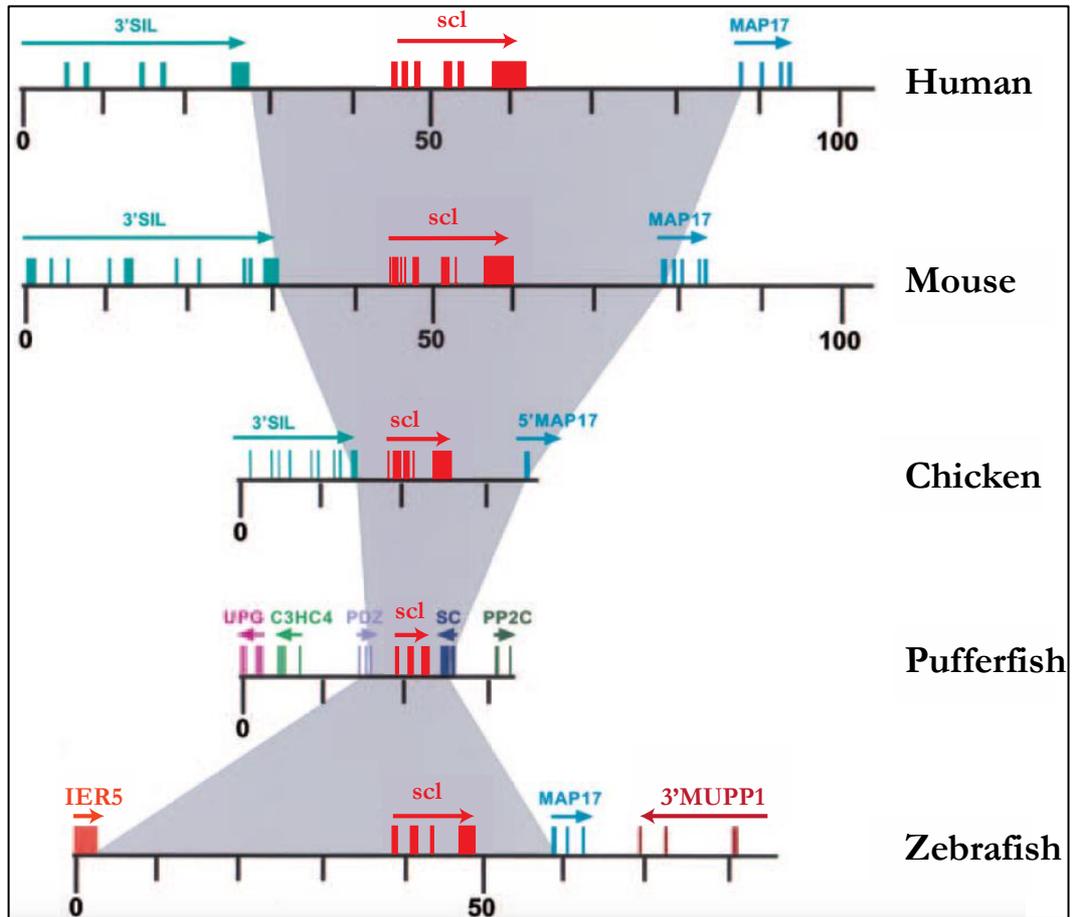
### 1.2.5. Regulation of Scl activity

#### Transcriptional Regulation

The comparison of genomic context within the *Scl* locus across 5 different vertebrate species has allowed identification of putative Scl regulatory elements as non-coding highly conserved regions<sup>155</sup>. In all five species the downstream locus is the same, though there is variation in the upstream neighbouring loci. In combination with expression data from these species it was concluded that the key regulatory regions for the Scl locus were also likely to be conserved<sup>17,99,104,155,156</sup>. Multiple sequence alignments identified areas of highly conserved sequence outside of the coding regions, including in the 3'UTR, the first non-coding exon and upstream elements (figure 1-5). The analysis of putative regulatory elements identified, using long-range comparative sequence analysis and phylogenetic foot-printing, binding sites motifs for GATA, SKN1 and YY1 transcription factors within these highly conserved regions<sup>155</sup>. Subgroups of T-ALL patients have been identified that display overexpression of specific haematopoietic genes, such as *runx1*, *myb*, *nkx3-1*, *erg* and *ets1*, suggesting regulatory interactions between these factors<sup>157-161</sup>. Sanda *et al.* confirmed Scl binding in proximity to these genes<sup>139</sup>. Sequences for E-box, GATA, Runx and ETS binding sites were over-represented in proximity to the identified Scl binding sites<sup>139</sup>. An autoregulatory loop involving Scl, GATA3 and Runx1 was identified with binding and occupancy sites within their own and each others' regulatory regions<sup>139,162-165</sup>. Conservation of key regulatory elements at the Scl locus

through five vertebrates, suggests similar regulatory circuitry may surround *Scl* across vertebrate taxa.

Figure 1-5- Synteny of the *Scl* locus across five vertebral species, by Gottgens *et al.* 2002<sup>155</sup>.



It has previously been shown that the *Scl* flanking sequences from the closely related species *Takifugu rubripes* are sufficient to recapitulate *Scl* expression in zebrafish following genetic ablation<sup>166</sup>

An element with novel regulatory function was identified in proximity of one of *Scl*'s enhancers in embryonic stem cell lines<sup>167</sup>. This motif carried active enhancer histone marks, but failed to promote expression in transgenic mice. This motif was discovered to boost activity of the proximal enhancer, possibly through competitive or constructive complex formation between the two regulatory elements<sup>167</sup>.

In human haematopoietic cell lines, histone and transcription factor ChIP has been used to describe the regulatory regions within the *SCL* locus. This identified enhancers, which when tested in reporter assays, suggested both positive and negative influences on *SCL* expression levels<sup>168</sup>. Combination of powerful techniques such as these has allowed describing *SCL* regulatory landscape at a previously unprecedented resolution and has begun to offer an insight into the gene regulatory control of this essential factor. However such techniques have been limited by their requirement for substantial input material and were feasible only in culture *in vitro* systems.

### **Transcript Degradation**

The 3'UTR conserved region contains sequences that upon transcription would likely form stem loop structures, which can provide binding sites for RNA-binding proteins<sup>169,170</sup>.

T-ALL has previously been related to disruption of normal microRNA (miRNA) networks<sup>171,172</sup> and a recent study has revealed that miRNA-mediated silencing of Scl may occur during normal T-cell development<sup>173</sup>. miRNA recognition sites were identified in the 3'UTR of human *SCL* transcripts and their disruption provided leukemic advantages. Over-expression of miRNAs predicted to bind to these sites caused *SCL* knockdown in cell culture<sup>173</sup>. A recent study has identified a number of miRNAs, potentially targeting *SCL* transcript, that were found down-regulated in T-ALL patient samples, suggesting their absence could account, at least in part, for ectopic activation of SCL in the this subgroup of T-ALL patients<sup>173</sup>.

### **Post Translational Regulation**

The Scl protein has been shown to undergo post-translational modification, in the form of phosphorylation<sup>174</sup>. The anti-apoptotic and pro-proliferative protein kinase, Akt, phosphorylates Scl at Thr90. This phosphorylation caused relief from Scl-

mediated repression at a target gene required for erythroid cytoskeletons<sup>174</sup>. Scl is subject to ubiquitination at its C-terminus, which tags the protein for proteasome-mediated degradation in response to Notch signalling<sup>175</sup>. Phosphorylation at serine 172 is required for association between Scl and the histone modifier, LSD1<sup>153</sup>. This is mediated by PKA, and PKA inhibitors have been shown to prevent Scl's interaction with LSD1 resulting in a de-repression of Scl target genes<sup>153</sup>.

### 1.3. The zebrafish model.

#### 1.3.1. Zebrafish general development

Zebrafish (*Danio rerio*) is a highly desirable model for the study of haematopoietic development in vertebrates. Embryos are externally fertilized, develop independently of the mother and are thus available for manipulation and observation starting from a single-cell stage. A single female zebrafish can produce several hundred eggs a week, lending statistical significance readily to any study carried out in this model. The generation time for zebrafish is comparable to that of mice, with animals reaching reproductive maturity within 2-3 months, making intergenerational studies feasible. Early development occurs at a faster rate than in other animals such as mice, which significantly increases the feasibility of lineage tracing studies and observation of developmental morphological changes in real time. Zebrafish embryos are optically transparent permitting non-invasive observation and high resolution imaging of internal developmental processes. These fish are relatively small, making animal facility costs for zebrafish more economical than for other model organisms such as mice, frogs or quail.

Invertebrate models, such as *Drosophila*, which have greatly contributed to our understanding of key early developmental processes, lack a complex circulatory system and show major immune system differences. The haematopoietic and

circulatory system of the zebrafish is complex and includes an adaptive immune system, both of which show high conservation to mammals.

Several key techniques were developed in zebrafish to aid in understanding of a wide range of biological and clinical processes, including development of the circulatory system. The high fecundity of zebrafish and relative ease of generating genomic modifications within the germ line, either in a non-targeted fashion using ENU, insertional mutagenesis, or gene trap approaches or in a targeted fashion via CRISPR/Cas9 genome engineering technology, makes zebrafish a highly tractable genetic system and one of the most useful vertebrate models. Like *Drosophila*, zebrafish has been used extensively in large-scale forward genetic screens, which have contributed a vast body of knowledge and afforded powerful insights into gene function during embryonic development<sup>176</sup>. The extensive studies of mutant zebrafish lines have provided us with key tools for investigating biochemical pathways involved in the regulation of disrupted processes.

Reverse genetic techniques are today predominantly employed to generate transgenic lines that express a gene of interest or mutant variant, often in combination with a fluorescent reporter or in a tissue-specific manner, using either classical or BAC-mediated transgenesis<sup>177</sup>. Genome editing approaches, particularly efficient in zebrafish, not only allow targeted gene interruption via non-homologous-end joining (NHEJ) repair, following CRISPR/Cas9-induced double stranded DNA breaks, but also enable precise mutational analysis, introduction of fluorescent protein, or targeted insertion of protein tags using homology directed repair (HDR) mechanisms<sup>178-180</sup>. Such approaches, yielding a large number of useful transgenic lines, targeting, tagging or modifying genes of interest or non-coding regulatory elements, offer great potential for deciphering *in vivo* molecular functions, while also permitting observation of cellular populations as they migrate or develop *in vivo*.

Transient ectopic expression of a gene of interest can be achieved through injection of mRNA encoding the gene into a single cell-staged zebrafish embryo. As the embryo develops the mRNA is taken up by the cells and translated resulting in ubiquitous ectopic expression. Conversely, endogenous gene expression can be specifically and transiently knocked-down in embryos through the use of morpholinos. Morpholinos are synthetic DNA analogues, which have standard DNA bases, bound not by deoxyribose, but with morpholino rings and as such represent foreign entity to the cells and are thus not recognised or degraded by cellular proteins<sup>181</sup>. These single stranded DNA oligomers bind to single stranded RNA in the cell and can disrupt gene function by either preventing the initiation of translation (when bound within the context of Kozak/first ATG) or alter the correct splicing of the pre-mRNA transcript in the nucleus (when targeted to important splice acceptor/donor sites) and hence lead to the generation of nonsense mutations or frame shifts. Historically this method has been widely used due to its relative ease and reproducibility, with many morpholinos recapitulating mutant phenotypes, but also significant number yielding off target effects and non-specific results.

The zebrafish has been used to model many human diseases (reviewed by Huiting *et al.*, North *et al.* and Cutler *et al.*<sup>182-184</sup>) and their ability to absorb drugs from their tank water makes them excellent candidates for larger scale chemical screens. A number of molecules identified through such high-throughput screens in zebrafish are currently used in the clinic to treat human conditions<sup>185,186</sup>.

### 1.3.2. Zebrafish model of haemangiogenesis.

As in mammals, zebrafish primitive and definitive waves of haematopoiesis also occur in different embryological locations. Zebrafish primitive haematopoiesis occurs within the embryo proper, in the lateral plate mesoderm instead of the extra-

embryonic yolk sac in amniotes (avian and mammalian models). The primitive wave of haematopoiesis in zebrafish involves the generation of erythrocytes and some primitive myeloid cells such as macrophages. The haemangiogenic lateral plate mesoderm arises as a pair of bilateral stripes of angioblasts shortly after gastrulation, which can be identified by the expression of key haematopoietic and vascular markers including *scl*<sup>187</sup>. Unlike in mammals, the development of primitive wave myeloid and erythroid lineages are spatially distinct, and take place in the anterior and posterior lateral mesoderm within the zebrafish embryo, respectively<sup>187,188</sup>. This affords us a significant advantage for investigating the molecular signalling pathways contributing to each of these lineages. Progenitors for myeloid and erythroid lineages share a significant portion of regulatory network interactions, as confirmed by the common expression of key haematopoietic factors such as *scl*, *etv2* and *lmo2*<sup>189</sup>. However distinct regulatory networks exist for each lineage as demonstrated by the zebrafish *spadetail* mutant which displays loss of the primitive erythroid compartment despite normal myelopoiesis in the anterior lateral plate mesoderm (ALM)<sup>44</sup>.

A transient wave of haematopoiesis follows the primitive wave and occurs in the posterior blood island (PBI). During this stage the first common haematopoietic progenitor arises, characterised by capacity to produce multiple blood cell types<sup>190</sup>. Despite the drop in *gata1* expression in the Intermediate Cell Mass (ICM), this key erythroid factor is maintained in the PBI, indicating a spatial shift in the site of erythropoiesis<sup>191</sup>. The common haematopoietic progenitor detected in the PBI at ~36hpf gives rise to both myeloid and erythroid lineages. However these progenitor cells lack self-renewal properties, and thus diminish in number as development progresses<sup>190</sup>. Shortly after the emergence of erythro-myeloid progenitors, the PBI undergoes massive remodelling to become the caudal haematopoietic tissue (CHT).

Definitive haematopoiesis generates HSCs capable of generating all blood lineages and of self-renewal, thus maintaining haematopoiesis throughout the adult life of the animal. In mammals definitive haematopoiesis originates in the aorta–gonad–mesonephros (AGM) region, specifically within the ventral wall of the dorsal aorta (DA), before moving to the foetal liver and onto the bone marrow. In zebrafish the origin of the definitive wave is in the equivalent location, within the ventral wall of the dorsal aorta (DA). HSCs then spread throughout the CHT, before migrating to the kidney, which remains the site of haematopoiesis throughout the life of the animal, in a similar fashion to bone marrow in mammals<sup>192,193</sup>.

*Runx1* gene, used as a marker of definitive haematopoiesis, is a key transcriptional regulator expressed in HSCs and required for the definitive wave of hematopoiesis<sup>194</sup>. *Runx1* expression commences shortly after the initiation of circulation, and is restricted to HSCs and endothelial cells of the ventral wall of the DA. As with the haemangioblasts, the close proximity of these endothelial and haematopoietic cells, as well as their resembling expression patterns have led several groups to propose the existence of a common progenitor for both of these lineages, called hemogenic endothelium (HE). This concept has been supported by single cell imaging studies in mice, which show that individual hemogenic endothelial cells can develop into haematopoietic cells<sup>195,196</sup>. Imaging of developing zebrafish observed that at ~24hpf expression of *scl*, *gata1* and *gata2* begins to reduce in the ICM and genes such as *runx1*, *c-myb* and *scl* start to be expressed within the AGM<sup>44,192,197</sup>. Use of antibodies against CD41, an integrin (cell surface marker) expressed on the surface of HSCs and absent from the endothelial cells of the DA, has permitted studies of HSC migration following their emergence from the ventral wall of the DA. In zebrafish, CD41 reporter lines have shown that HSCs move to the CHT, developing thymus and colonize the kidney by 2-3 dpf<sup>197</sup>. By 6dpf the kidney becomes the key site of

haematopoiesis for all blood lineages as blood development diminishes in the CHT<sup>197</sup>. Expression of *e1-globin* and *gata1* in the CHT at 3.5dpf indicates the initiation of definitive erythropoiesis, which gradually moves to the kidney by 5dpf<sup>198,199</sup>. Switching between embryonic globins and adult globins that contribute to the haemoglobin tetramer, a process conserved within higher vertebrates, occurs after 15dpf.

The initiation of *l-plastin* expression in the CHT indicates the initiation of definitive myelopoiesis at 3dpf, but this process moves to the kidney by 4dpf, which remains the site of HSC development into all blood lineages throughout adulthood<sup>199</sup>.

Lymphopoiesis is responsible for generating the cells required for adaptive immunity, and signalling pathways involved in this process have been conserved between zebrafish and mammals. T and B cells are produced and require RAG mediated recombination to generate mature antigen receptors. Common lymphoid progenitors arise from HSCs in the kidney, and while B-cell progenitors remain there to maturity, the T-cell progenitors migrate to the thymus to complete maturation<sup>200</sup>. Innate and adaptive immune development is temporally distinct in zebrafish, with the mature T-lymphocytes only arising in the thymic epithelia at 7dpf<sup>200</sup>. Similarly to in mammals, zebrafish thrombocytes are capable of mediating clot formation in response to injury. The first thrombocytes to form are in the CHT at 2dpf, but later in development are also formed in the kidney<sup>201</sup>.

### 1.3.3. Developmental origins of the *scf*<sup>+</sup> population

*scf* knockdown in zebrafish ablates primitive and definitive haematopoietic cell lineages and causes disruption of angiogenic patterning<sup>202</sup>. Ectopic expression causes expansion of haemangioblast population and subsequently, the enlargement of

erythropoietic compartment, leading to malformation of the circulatory system of the fish<sup>18,203</sup>.

Davidson *et al.* 2003 detail the initiation of *scf* expression in zebrafish and show it onsets at 2-somite stage (2ss) in putative HSC and angioblasts<sup>187</sup>. In 4ss zebrafish embryos, *scf* is expressed as two bilateral stripes both in the anterior and posterior lateral mesoderm (ALM and PLM)<sup>203</sup>. *scf* expression in these domains is accompanied by the expression of other key haemangiogenic factors such as *flk1*, *lmo2* and *gata2*<sup>189</sup>. As in other vertebrates, such expression data was used to support the existence of the common vascular and haematopoietic precursor, the haemangioblast, in zebrafish as well. During development, from the 14-18ss, the PLM stripes of *scf* expression extend anteriorly and posteriorly. By 20ss all *scf*-expressing PLM cells migrate medially and converge at the midline and by 20ss to form the main primitive erythroid tissue of the embryo, the Intermediate Cell Mass (ICM)<sup>198</sup>, which still expresses *scf*. Further evidence suggesting existence of haemangioblasts comes from fate-mapping experiments of single cells in the ventral mesoderm of zebrafish embryos, which showed that individual cells can give rise to both haematopoietic and vascular cells in the ICM<sup>204</sup>. Other vascular and haematopoietic markers are co-expressed with *scf* by the cells in the ICM.

Erythroid markers such as *gata1* are expressed in the PLM as early as the 4-5ss. These *gata1*<sup>+</sup> cells develop into the first circulating erythrocytes<sup>44,188</sup>. This strongly suggests that haematopoietic lineage fates may be specified as early as 4ss, but they have proved highly challenging to investigate or describe due to their inherent transient nature and small cell number.

At ~24hpf circulation commences and around 300 pro-erythroblasts from the ICM enter circulation and continue erythroid maturation and proliferation. Transcription factors such as *runx1* and *fli1a* have been observed to reduce their expression levels,

indicating the cessation of primitive erythropoiesis at this time (24hpf). These primitive wave erythroid cells provide the complete oxygen transport function for the embryo, during the first 4 days of zebrafish development.

Angioblasts of the anterior lateral plate mesoderm (ALM) are reported to display a similar expression profile as their posterior counterparts, but lack *gata1* expression<sup>191</sup>. Instead *pu.1/spi1b*, a myeloid specific transcription factor, is detected in *scf*<sup>+</sup> ALM cells, as early as at 3ss<sup>205</sup>. As development progresses these cells disperse across the yolk and migrate rostrally, to ultimately differentiate into granulocytes and macrophages<sup>191</sup>. Molecular studies have identified two myeloid populations within the anterior *scf*<sup>+</sup> population<sup>191</sup>. *l-plastin* is an actin binding protein specifically expressed by all myeloid cells<sup>191</sup>. A subset of *l-plastin*<sup>+</sup> cells also express *mpo*, a peroxidase specifically transcribed in granulocytes. Macrophages and monocytes are identified as *l-plastin*<sup>+</sup>/*mpo*<sup>-</sup> cells of the ALM, while *l-plastin*<sup>+</sup>/*mpo*<sup>+</sup> cells are defined as granulocytic precursors<sup>191</sup>. Innate immunity develops rapidly in zebrafish, as these primitive macrophages have been reported as active as early as 26hpf, and while their full distribution through the embryo is established by 28hpf<sup>206,207</sup>.

Macrophages are the most functionally and spatially diverse cells of the haematopoietic system and play key roles in immunity, repair, homeostasis and development<sup>208-210</sup>. In mice the macrophage lineage originates from the yolk sac and foetal liver<sup>210-212</sup>. All macrophage subtypes have been shown to express Cell Stimulating Factor 1 Receptor (CSF1R), knock out of CSF1R causes severe depletion of macrophages in many tissues including loss of microglia (macrophages of the brain) and Langerhans cells<sup>213-215</sup>. However loss of *Spi1b* in mice has been shown to result in the complete loss of all macrophage lineages (plus causes depletion of B-cells)<sup>216</sup>, suggesting that *Spi1b* functions earlier in haematopoietic lineage development than CSF1R.

Both *spi1b* and *gata1* are expressed in the ICM, where they counter-regulate each other, producing a balance between primitive erythropoiesis and myelopoiesis in the posterior of the embryo<sup>217,218</sup>. Any disruption of either transcription factor results in up-regulation of one lineage with concurrent restriction of the other.

#### 1.3.4. Zebrafish *scf* isoforms

Multiple isoforms of *Scf* have been detected in both murine and human culture systems, indicating functional diversity<sup>68,103,108,219-221</sup>. Zebrafish embryos express two isoforms of *scf* that differ due to an N-terminal truncation, arising from a secondary transcriptional start site<sup>222</sup>. The originally described gene is *scf- $\alpha$* , with its transcription starting at the first transcriptional start site and produces a full-length protein. Transcription from a second alternative promoter site situated in the middle of the exon 2 of the *scf* locus produces *scf- $\beta$*  transcripts that upon translation produce a truncated protein that lacks the first 118 N-terminal amino acids. The expression patterns of the two isoforms overlap, with the *scf- $\beta$*  (truncated) isoform expressed first at 1-2ss, encompassing the ALM and PLM and later being detected in the ventral wall of the DA, where it is co-expressed with *c-myb*, indicating that *scf- $\beta$*  characterises definitive HSCs. *scf- $\alpha$*  expression begins at 4ss in a subset of *scf- $\beta$*  expressing cells within the posterior region of the ALM and in the majority of the PLM. By 26hpf expression of *scf- $\alpha$*  was undetectable in the ventral wall of the DA<sup>222</sup>. The two isoforms have been shown to be functionally redundant for the initiation of primitive haematopoiesis, however the isoforms differ in their contribution to primitive erythropoiesis. *Scf* isoform-specific morpholino knockdown experiments revealed that following *scf- $\beta$*  knockdown red blood cells were lost by 3 dpf<sup>222</sup>. Knockdown of *scf- $\alpha$*  isoform also resulted in anaemia, however erythroid cells were normal up to 3 dpf, but then have significantly decreased in numbers by 5 dpf. This

indicated that both isoforms were crucial for erythroid maturation but required at different stages.

*scf*- $\beta$  was also shown to be essential for definitive haematopoiesis, as in 30hpf *scf*- $\beta$  morphants, *runx1* and *c-myb* expression was lost in the ventral wall of the DA<sup>222</sup>. Zhen *et al.* have demonstrated that *scf*- $\beta$  marks the hemogenic endothelium (HE), while *scf*- $\alpha$  expression onsets only as HSCs emerge from the DA wall<sup>223</sup>. In *runx1* knockdowns, endothelial to HSC transition (EHT) is defective and cells attempting to make this transition undergo apoptosis<sup>224</sup>. Specific knockdown of *scf*- $\beta$  resulted in HE cells failing to form and a subset of cells in the ventral wall of the DA were observed to undergo apoptosis before any morphological changes associated with EHT were observed<sup>223</sup>. In *scf*- $\alpha$  morphants, EHT occurred normally but *c-myb* expression dropped dramatically between 2.5-3 dpf. This was discovered to be due to apoptosis of HSCs a few hours after undergoing EHT, suggesting a role for *scf*- $\alpha$  in the maintenance of HSCs. The role of both *scf* isoforms in definitive haematopoiesis was demonstrated to be downstream of *etv2*, however only *scf*- $\alpha$  was capable of rescuing both angioblast and HSC defects in *etv2* morphants<sup>225</sup>. The expression of *scf* isoforms was also shown to be downstream of VEGF-signalling during definitive haematopoiesis in zebrafish, though the use of a VEGFR inhibitor, SU5416, which in mice specifically inhibits Flk-1/KDR activity<sup>226</sup>. Overexpression of either of the isoforms can partially rescue *runx1* expression in VEGF signalling-deficient embryos<sup>225</sup>.

Protein stability was also noted to be significantly different between the two isoforms, with *scf*- $\beta$  undergoing rapid degradation at the protein level rather than at the RNA level<sup>222</sup>.

## 1.4. Overview

In this project I have investigated the dynamic nature of early haematopoietic and vascular progenitors in zebrafish, identified through their expression of the blood and vascular regulator, *scf*.

The initial aims of the project were to describe how the posterior erythrogenic population develops during primitive haematopoiesis and to identify early differences between the anterior and posterior primitive haematopoietic progenitor populations.

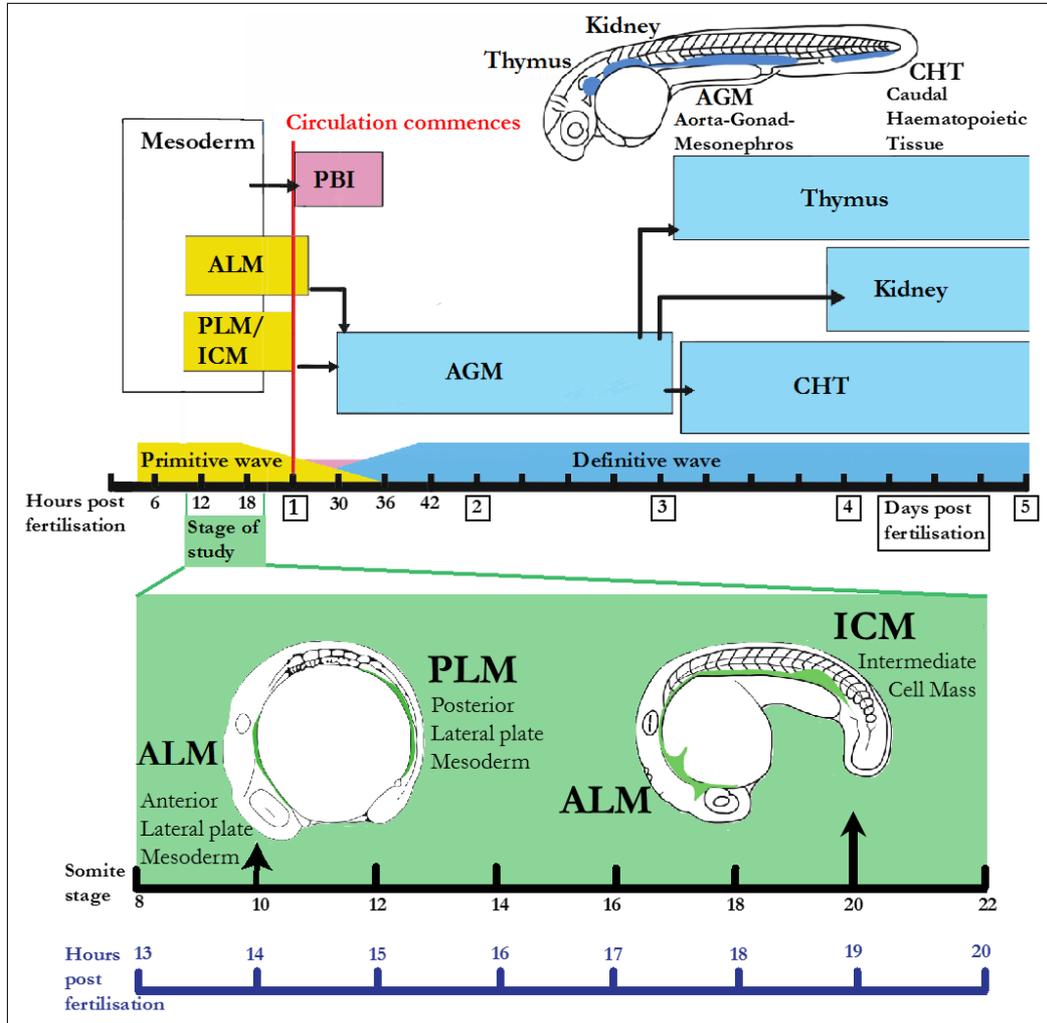
Previous studies have identified *scf* expression specifically within early haematopoietic progenitor populations, thus *scf* was used as a molecular marker for the populations of interest for this project.

To identify any changes in early haematopoietic progenitor activity, it is necessary to compare different biological contexts, which I have isolated from early stages of development in zebrafish embryos. I have focussed on the earliest stages of *scf* expression (10 and 20ss) and performed both spatial and temporal comparisons (anterior and posterior embryonic regions), to investigate how *scf* regulatory circuitry varies with embryonic development. I have generated a tagged transgenic *scf* reporter line in zebrafish and characterised its expression pattern throughout early development. This line spikes the endogenous Scl protein population with a tagged transgenic Scl protein, permitting isolation of the tagged proteins plus bound DNA and protein partners in future studies. In addition, a fluorescent reporter is also expressed, which labels the *scf*<sup>+</sup> cells *in vivo* and enables specific cell isolation through fluorescent activated cell sorting (FACS).

Temporal comparisons have been performed using the posterior *scf*<sup>+</sup> population, which has been previously been determined as the main site of erythropoiesis. In this investigation I have described the transcriptional and chromatin accessibility

changes genome-wide, associated with the *scf* population progressing from haemangiogenic precursors to maturing primitive erythroid cells.

**Figure 1-6 Schematic of haematopoietic development in the zebrafish highlighting the contexts of investigation in this project.**



Profiles of the 10ss anterior *sc<sup>l</sup>*<sup>+</sup> population revealed significant diversity in associated biological functions. By applying the profiling techniques used to compare the early progenitor populations at single cell resolution, I was able to investigate the early cellular diversity in the anterior *sc<sup>l</sup>*<sup>+</sup> populations. I used *in vivo* confocal microscopy and single cell RNA sequencing to determine the heterogeneity of this population and to gather the necessary data to describe ensuing cell subpopulations.

In summary this project's key aims were to

- A- Describe the development of primitive erythropoietic progenitors
- B- Investigate early spatial differences in primitive haematopoietic progenitor populations.
- C- Assess the cellular diversity of early anterior haematopoietic progenitors.

Each of these aims would be performed within the context of the developing zebrafish embryo. These investigations were planned to add true vertebrate *in vivo* knowledge of crucial developmental populations that have been previously only accessible through cell culture and primary culture modelling.

## 2. Materials and Methods

### 2.1. Molecular techniques

Primer sequences are detailed in Table 2-2 and solution recipes are listed at the end of each sub-chapter.

#### 2.1.1. BAC transgenesis

##### **Generation of *scI* BAC material.**

BAC zC104B7 (Chori211-104B7) was ordered from Chori BAC PAC facility, this contained chromosomal regions 22:16713150-22:16713954 plus a BAC backbone, which include a chloramphenicol resistance cassette. Upon delivery of the zC104B7 bacterial stub, a swab was plated on agar plates with chloramphenicol at 25 ug/ml and grown for 20 hours at 30°C. Single colonies were used to seed starter cultures containing 12.5 ug/ml chloramphenicol in LB-broth, which following 20 hours of growth at 30°C were used for DNA extraction. Bacteria were pelleted by centrifugation at 5000g at 4 for 20 minutes. Pellets were resuspended in 250ul buffer P1 from QiaPrep spin miniprep kit (Qiagen), 250ul buffer P2 was added and the mix was incubated at room temperature (RT) for 4 minutes. 250ul of precool buffer P3 was added followed by 10-minute incubation on ice. This lysate was cleared by centrifugation at 4°C at 16500g for 10 minutes. 750ul of isopropanol was added to the supernatant and then incubated on ice for a further 10 minutes. BAC DNA was pelleted by centrifugation at 16500g for 15 minutes, then washed with 70% ethanol. BAC DNA was finally resuspended in 50ul of nuclease free H<sub>2</sub>O and incubated at 50°C for 30 minutes to aid full resuspension.

## Cloning N-terminally tagged *scl* template vector for BAC transgenesis.

To confirm the recombination of the insert into the BAC, I used a *frt-kan-frt* (*fkf*) cassette. This is a kanamycin resistance cassette that would provide additional antibiotic selection for the insert, flanked by two *frt* sites. *Frt* sites are recognized by flippase, which is employed at the final stage of BAC transgenesis to remove this antibiotic resistance from the BAC. Removal of the kanamycin resistance cassette ensures that transgenesis does not result in additional antibiotic resistance within our zebrafish stocks.

A *pGEM-Teasy:citrine-fkf-sv40pA* cassette was available in one of my host laboratories. This was linearized with *SacII* (NEB), blunted with T4 polymerase then further digested with *BsrGI* (NEB) to generate two fragments- 3.1kb and 1.7kb. *NC1:avi-tev-flag-scl-2A-egfp* was digested with *NcoI*, blunted with T4 polymerase and then digested with *BsrGI* yielding two fragments at 7.5 and 1.8 kb. The *pGEM-Teasy* vector (3.1kb) and N-terminally-tagged-*scl* (1.8kb) fragments were gel extracted and ligated together at a 5:1 insert to vector ratio with T4 ligase. The ligation mixes were electroporated into DH10B and bacteria expanded as before were used to produce *pGEM-Teasy:avi-tev-flag-scl-2A-citrine-fkf-sv40pA* plasmid DNA using a Qiagen miniprep kit.

## Cloning C-terminally tagged *scl* template vector for BAC transgenesis.

*NC1:avi-tev-flag-scl-flag-tev-avi-2A-egfp* was digested with *NcoI* and *pGEM-Teasy:citrine-fkf-sv40pA* was digested with *SacII* (NEB), both were then blunted with T4 polymerase treatment. Each of the linearized constructs were then digested with *BsrGI* (NEB) then the 8kb *pGEM-Teasy* vector band and the *scl-flag-tev-avi-2A* 3kb band, were gel extracted. A T4 ligation was set up at a 3:1 ratio of insert to vector.

Transformation, bacterial colony amplification and DNA purification were carried out as previously described, to give *pGEM-Teasy:scl-flag-tev-avi-2A-citrine-fkxf-sv40pA* DNA sample.

### **Generation of electrocompetant SW105 cells**

SW105 cells were used for BAC transgenesis. This strain contains a heat inducible recombinase and an arabinose inducible flippase, which were used to insert the *tol2* cassette and to flip out the kanamycin resistance cassette, respectively.

100ml SW105 culture was grown to an optical density (OD) of 0.5 in LB broth with chloramphenicol at 12.5 ug/ml at 30°C. The culture was divided into two 50ml cultures. One culture is incubated at 42°C with agitation for 15 minutes- this activates expression of recombination genes and was referred to as “experimental”, while the other culture continues to be incubated at 30°C. Both cultures were then incubated on ice for 5 minutes before centrifugation at 2000g for 15 minutes at 4°C. Supernatant was discarded, cells washed in ice cold H<sub>2</sub>O and centrifugation repeated. Again supernatant was discarded and pellets resuspended in ice cold 10% glycerol. After a further centrifugation the supernatant was fully removed and the pellet resuspended in 500ul GYT medium, and experimental and control SW105 50ul aliquots were stored at -80°C.

### **Generation of recombination cassettes for BAC transgenesis.**

Recombination cassettes were amplified from their appropriate pGEM vectors using the PCR conditions below. *pGEM-Teasy:scl-flag-tev-avi-2A-citrine-fkxf-sv40pA* was amplified with primers Recomb\_SA\_F1 and Recomb\_R1 or Recomb\_Stop\_F1 and Recomb\_Stop\_R1 to yield a 2<sup>nd</sup> exon insert or stop codon insert, respectively. *pGEM-Teasy:avi-tev-flag-scl-2A-citrine-fkxf-sv40pA* was amplified with primers Recomb\_AS\_F1 and Recomb\_R1 to produce a 2<sup>nd</sup> exon insert.

0.3ul 500mM pGEM vector containing tagged Scl

1ul	10mM	Forward recombination primer (F1s)
1ul	10mM	Reverse recombination primer (R1s)
25ul	2x	PfuUltra II Hotstart mix
22.7ul		H <sub>2</sub> O

PCR program- repeat steps 2-4 x25

- 1) 95°C for 5 minutes
- 2) 95°C for 30 sec
- 3) 60°C for 30 sec
- 4) 72°C for 2:30 minutes
- 5) 72°C for 5 minutes
- 6) 4°C on hold.

The resulting PCR product was digested with DpnI (Roche) to remove plasmid DNA template, then gel extracted. 500ng of the resulting cassette was then electroporated into both experiment and control SW105 cells. Following electroporation, cells were incubated for 4 hours in SOC medium at 32°C with agitation then pelleted by centrifugation at 2000g for 15 minutes. Cells were then transferred to agar plates containing chloramphenicol at 12.5 ug/ml and kanamycin at 50ug/ml and incubated for 24 hours at 32°C. Individual colonies were picked and streaked on ampicillin only and chloramphenicol/kanamycin plates and incubated for 20 hours at 32°C. Colonies that grew on chloramphenicol/kanamycin and not on ampicillin only contained successfully recombined BAC and did not contain the original *pGEM-Teasy* plasmid. A single colony was grown in 25ml LB broth with 12.5 ug/ml chloramphenicol until it had reached an OD of 0.5. 25ul of 10% arabinose was added to the culture and incubated for a further two hours at 32°C to induce the expression of the flippase genes. An aliquot of the culture was plated on kanamycin only plates and incubated for 20 hours at 32°C. Individual colonies were picked and

streaked on chloramphenicol/kanamycin plates and kanamycin only and incubated for 20 hours at 32°C. Colonies that grew on kanamycin only and not on chloramphenicol/kanamycin contained the *scI* locus BAC that had successfully “flipped out” of the kanamycin cassette. Individual colonies were used to prepare experimental and control electrocompetant SW105 cells as described above.

The *tol2* cassette was amplified from the *pCS2:tol2* vector as detailed below.

1ul	5ug/ml	Tol2 vector
1ul	10mM	ptarbac loxp 5' primer
1ul	10mM	ptarbac loxp 3' primer
25ul	2x	PfuUltra II Hotstart mix
22ul		H <sub>2</sub> O

PCR program- repeat steps 2-4 x30

- 1) 95°C for 5 minutes
- 2) 95°C for 30 sec
- 3) 59.9°C for 30 sec
- 4) 72°C for 2:30 minutes
- 5) 72°C for 7 minutes
- 6) 4°C on hold.

The PCR product was treated with DpnI overnight and the *tol2* cassette was gel extracted. The *tol2* cassette was electroporated into experimental and control SW105 cells produced from the zC104B7 recombination with tagged *scI* constructs following kanamycin flip out. Cells were incubated for 4 hours in SOC medium at 32°C with agitation then pelleted as before by centrifugation at 2000g for 15 minutes. Cells were then transferred to agar plates containing chloramphenicol/ampicillin plates, which were then incubated for 24 hours at 32°C. Colonies that grew on chloramphenicol/ampicillin contained successfully recombined BAC, without a

kanamycin resistance cassette but with a *tol2* cassette that would mediate integration into the zebrafish genome. mg quantities of the final BACS were produced using the PureLink HiPure Plasmid Midiprep Kit (Invitrogen).

### **gDNA extraction**

F<sub>1</sub> clutches of 50 embryos or greater were grown to the 18hpf, dechorionated and used for DNA extraction with PureLink Genomic DNA Mini Kit (LifeTech), following kit instructions.

### **PCR screening**

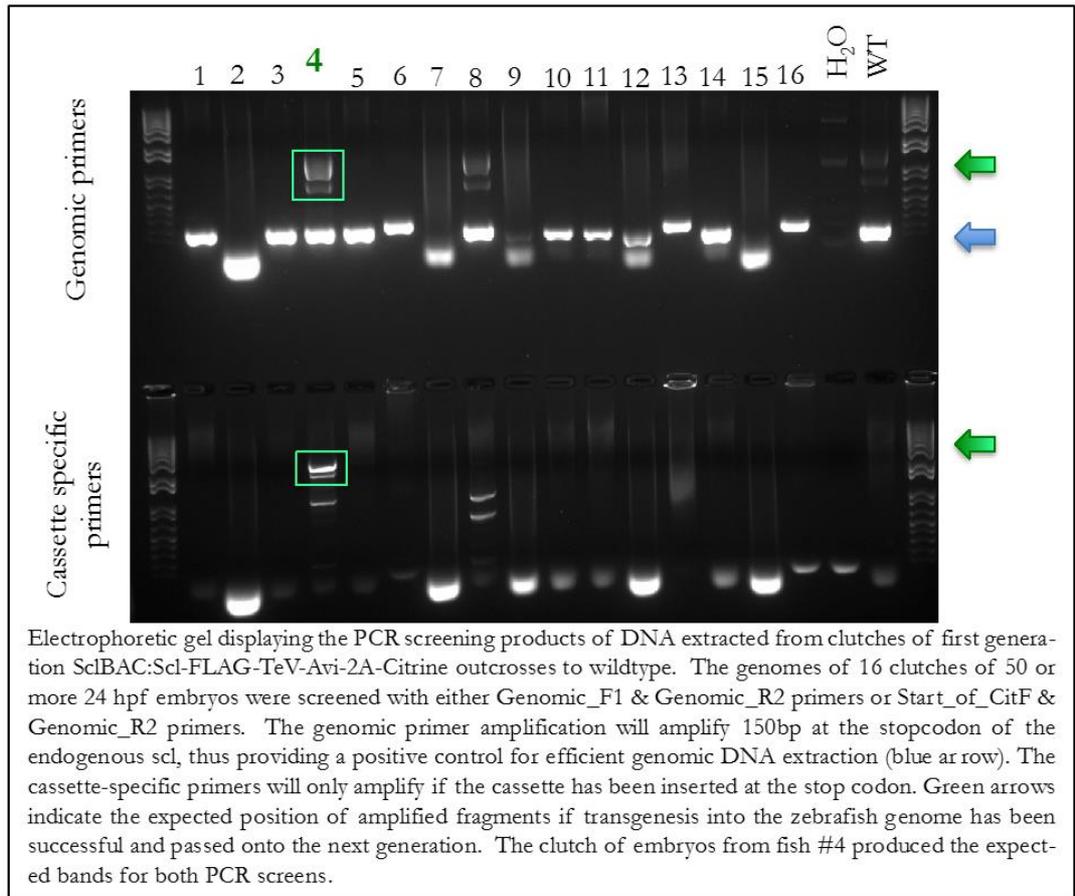
Screening for transgene incorporation into the genome of F<sub>1</sub> clutches was carried out using the following PCR conditions. The results are shown in figure 2-1.

0.3ul	10mM	Start_of_CitF or Genomic_F1
0.3ul	10mM	Genomic_R2
0.3ul	10mM	dNTP mix
0.15ul		HotStart Taq Polymerase (NEB)
1.5ul	10x	HotStart Taq Polymerase Buffer (NEB)
1ul		Extracted gDNA
11.45ul		H <sub>2</sub> O

#### PCR program- repeat steps 2-4 x35

- 1) 95°C for 5 minutes
- 2) 95°C for 30 sec
- 3) 55°C for 30 sec
- 4) 72°C for 4 minutes
- 5) 72°C for 7 minutes
- 6) 4°C on hold.

Figure 2-1 PCR screening for F<sub>0</sub>s with germline transmission of the *scl-citrine* transgene.



### ST<sub>low</sub>E

1ml 5M NaCl

1ml 1M Tris pH8.0

20ul 0.5M EDTA

97.98mls H<sub>2</sub>O

## 2.2. Fish protocols and husbandry

### mRNA Injections

mRNA was injected into single-cell stage zebrafish embryos, using the Picospritzer II microinjector (Parker Instrumentation). Embryos were incubated in E3 medium (Westerfield, 1993<sup>227</sup>) at 28.5°C until desired stage was reached.

## DNA Injections

DNA in plasmid form was injected at 150pg per embryo, at the single cell stage. BAC injection amount for *scBAC:sc-flag-tev-avi-2A-citrine-sv40pA* was 100pg, with 100pg of *tol2* mRNA, per embryo. Injection volume varied but never exceeded 2nl.

### E3 medium 60x stock

17.2g NaCl

0.76g KCl

2.9g CaCl<sub>2</sub>·2H<sub>2</sub>O

4.9g MgSO<sub>4</sub>·7H<sub>2</sub>O

## 2.3. Expression analysis

### 2.3.1. Transcript detection

#### Labelled probe production

RNA was isolated from a clutch of 10ss wildtype embryos, using RNAqueous MicroKit (Ambion) following kit instructions, including DNaseI treatment. cDNA was produced using Superscript III (Invitrogen) and purified using cDNA filter cartridges (LifeTech). Resulting cDNA was used as template for PCR amplification, using the protocol detailed in the “PCR screening” section above, using the 5’ and 3’ primers detailed in table 2-2 as appropriate. The correct PCR product was gel extracted and 5ug used for *in vitro* transcription with the T7 RNA polymerase kit (Promega) and DIG RNA labelling kit (Sigma) following kit instructions. The resulting RNA was treated with Turbo DNase (LifeTech) for 15 minutes at 37°C before use.

### ***In situ hybridization***

Embryos were fixed for 1hr at RT in 4% PFA. Fix was washed out with PBS-T then embryos were dehydrated and rehydrated using a methanol: PBST gradient. Embryos are washed in PBS-T, permeabilized with 1% $H_2O_2$  for 10 minutes on ice, then brought to Hybe-50. Labelled probes were applied to embryos in hybe-50 then incubated overnight at 65°C. Embryos were washed in Hybe-50 at 65°C then in MABT at RT. Embryos were blocked with 20% sheep serum 2% Boehringer Blocking Reagent (BBR) (Sigma) for 1 hour at RT. Embryos were then incubated with anti-dig-AP antibodies <sup>184</sup> overnight at 4°C. Further MABT were carried out before developing colour using BN purple reagent (Sigma).

### **RT-PCR analysis**

RNA was isolated using the RNAqueous MicroKit (Ambion) following kit instructions, including DNaseI treatment. 5ug of isolated RNA was used for reverse transcription using the reaction mixes detailed below.

1ul 50uM oligodT primers

5ug isolated RNA

1ul 10mM dNTP mix

Up to 13ul with  $H_2O$

This annealing mix was heated to 65°C for 5 minutes then incubated on ice for addition of the following,

4ul 5x First strand synthesis buffer (Invitrogen)

1ul 0.1M DTT

1ul RNaseOUT Recombinant RNase inhibitor (Invitrogen)

1ul Superscript III reverse transcriptase (Invitrogen).

This reverse transcription mix was accompanied by a negative control, prepared in the same manner except omitting the reverse transcriptase. Both reactions were incubated at 25°C for 5 minutes then at 50°C for 1 hour.

cDNA produced was amplified using the PCR conditions detailed below. 100ng of *pGEM:scl-flag-tev-avi-2A-citrine-3'UTR* was used as a positive control and these PCR conditions would yield a comparative fragment to fully spliced transgenic mRNA.

0.3ul	10mM	RT-PCR-Fwd
0.3ul	10mM	Mid_Cit_R
0.3ul	10mM	dNTP mix
0.15ul		HotStart Taq Polymerase (NEB)
1.5ul	10x	HotStart Taq Polymerase Buffer (NEB)
1ul	100mM	cDNA
11.45ul		H <sub>2</sub> O

PCR program- repeat steps 2-4 x35

- 1) 95°C for 5 minutes
- 2) 95°C for 30 sec
- 3) 55°C for 30 sec
- 4) 72°C for 2 minutes
- 5) 72°C for 5 minutes
- 6) 4°C on hold.

PCR products were then run out on a 1% agarose electrophoretic gel.

MABT (500mls) pH7.5

50mls 1M Maleic acid

15mls 5M NaCl

0.5ml 100% Tween-20

Hybe-50 (500mls)

250mls 100% Formamide

32.5ml 20x SSC-DEPC treated

5ml 0.5M EDTA

5ml 20mg/ml tRNA

1ml 100% Tween-20

25ml 10% CHAPS

50mg Heparin

Bring up to 500ml with H<sub>2</sub>O

KTBT (1l)

50ml 1M Tris-HCl pH7.5

30ml 5M NaCl

10ml 1M KCl

20ml 10% Tween-20

90mls H<sub>2</sub>O

## 2.4. Sample Preparation.

### **Dissociation of Embryonic Samples**

Initial embryo number varied between 40 and 250 embryos, depending on clutch size and survival. Embryos were dechorionated, fluorescently sorted and then bisected as detailed in Figure 3-10. Embryo halves were dissociated by incubation with 20mg/mg collagenase in trypsin solution for 9 minutes at 30°C. Dissociation is

halted by addition of Hank's solution and remaining cell aggregates dissociated by gentle pipetting. Centrifugation at 500g for 15 minutes pelleted the cells and supernatant was discarded. The cell pellet was washed further in Hank's solution, passed through a 40um cell strainer and centrifuged again. The final cell pellet was resuspended in 200-400ul of Hank's Solution.

### **FACS**

FACS was carried out using a FACS AriaIII (BD Biosciences) within the in-house cell sorting facility, using a 100 micron nozzle with precision calibrated for purity. FACS enabled separation of cells based on their size, granularity and citrine fluorescence after calibration against WT cell samples.

### **RNA isolation**

RNA extraction was carried out on samples containing a minimum of 5000 citrine+ FACS sorted cells, immediately after sorting using the RNAqueous MicroKit (Ambion) following kit instructions, including DNaseI treatment.

### **RNA quantification**

RNA samples were quantified using a Bioanalyzer (Agilent) using the RNA 6000 Pico reagents (Agilent). Samples were required to show clear peaks for the 5.8 and 18s rRNAs and an RNA integrity number (RIN) of 7, to proceed with sequencing. 5ng of each sample was sent off for sequencing sample preparations with a Kapa HIFI ready mix.

### **ATAC**

Immediately following FACS *citrine*<sup>+</sup> samples of 15000 cells or greater, were spun down at 500g for 5 min at 4°C and supernatant discarded. Cells were washed once with ice cold PBS and then spun down again at 500g for 5 min at 4°C. The cell pellet was resuspended in 50ul ATAC lysis buffer and then spun down at 500g for 10min at 4°C and supernatant discarded. For 10,000-50,000 cells the pellet was

resuspended in the following transposition mix and incubated at 37°C for 30 minutes.

10ul 2x TD Buffer (Illumina)

1ul Tn5 transposase (Illumina)

9ul H<sub>2</sub>O

This reaction was purified using the PCR purification MinElute kit (Qiagen) following the kit instructions and eluting in 10ul elution buffer. Transposed DNA was amplified using the following PCR reactions.

20ul Transposition mix

2.5ul Customized Nextera PCR Primer 1 universal AD2.1 (Illumina)

2.5ul Customized Nextera PCR Primer (varied to allow sample pooling) (Illumina)

25ul NEBNext High Fidelity 2x PCR master mix (NEB)

PCR program- repeat steps 3-5 x11

1) 72°C for 5 minutes

2) 98°C for 30 sec

3) 98°C for 10sec

4) 63°C for 30 sec

5) 72°C for 1 minute

6) 4°C on hold.

The resulting ATAC library was purified with the PCR Purification MinElute kit again and eluted with 20ul elution buffer. Following this Ampure purification was carried out, with incubation of the library with 20ul of Agencourt Ampure beads (Beckman Coulter) at RT for 5 minutes. Using a magnetic stand beads were retained and supernatant removed, permitting washing of beads with 80% EtOH twice. DNA was eluted from the beads with 20ul 1mM Tris HCl pH8.0 and 1mM EDTA.

### ATAC library quantification,

ATAC libraries were analysed and quantified using the TapeStation machine with the D1000 High sensitivity and reagents (Agilent), following kit instructions. QBit fluorometric quantitation (ThermoFisher) was used to precisely measure the concentration of the ATAC library. 0.08pmoles of each ATAC library was used for high-throughput sequencing as detailed for RNA samples above.

#### Hank's Solution

2.5ml	10x	HBSS
62.5mg		BSA
0.25ml	1M	HEPES pH8.0
22.25ml		H <sub>2</sub> O

#### ATAC Lysis Buffer

10ul	1M	Tris HCl pH7.5
2ul	5M	NaCl
10ul	10%	NP-40
3ul	1M	MgCl <sub>2</sub>
975ul		H <sub>2</sub> O

## 2.5. Bioinformatic analysis.

### 2.5.1. Generation of *scf* specific transcriptomic data

75bp paired-end reads were produced on a NextSeq Illumina platform with a NextSeq™ 500 High Output Kit (150 cycles) (Illumina). FastQC<sup>228</sup> was used to check the quality of the reads. This package identified over-represented sequences relating to the poly-A tail sequence, TruSeq adapter sequences and a sequencing primer. To account for this reads were trimmed using Scythe (v0.994Beta)<sup>229</sup>.

Trimmed reads were mapped to the zv10 genome using STAR<sup>230</sup>. Expression level was calculated as FPKM values, which accounts for transcript length and for the total sequencing depth of the sample<sup>231</sup>. Conversion to FPKM values permits meaningful comparison between samples. DESeq used for differential expression analysis<sup>232</sup> as this package took into account the replicates of samples thus could compare biological variation within replicates to variation between samples. Expression was defined as any gene showing expression above or including 2 FPKM, this has previously been used as a cut-off for expression 2 in genome-wide population analyses<sup>233</sup>.

### 2.5.2. ATAC-Seq

40bp paired-end reads were produced on a NextSeq Illumina platform with a NextSeq™ 500 High Output Kit (150 cycles) (Illumina). Read quality was assessed using FastQC<sup>228</sup> and mapped to the zv10 genome using bowtie(v1.1.2)<sup>234</sup>. Duplicate reads caused by over-clustering and PCR represented 0.33% of mapped reads and were removed using PicardTools (v.1.83) MarkDuplicate<sup>235</sup>. DeepTools (v.2.2.2.) was used to assess the quality of mapped reads<sup>236</sup> and SamTools was used to process sam files<sup>237</sup>. Bam files were converted to paired-end bed files using BedTools (v2.25.0) with the bam2bed-bedpe package<sup>238</sup>. D. Gavriouckina carried out further mapping, filtering and peak calling. Buenrostro *et al.* determined that a 4 and 5 bp shift of mapping location yielded more accurate peaks for open chromatin location<sup>239</sup>. This group also identified a high percent of reads mapped to the mitochondrial genome<sup>239</sup>. Reads were mapped to the mitochondrial genome to identify the mitochondrial DNA contribution, which represented under 10% of reads following duplicate removal by PicardTools, and reads were adjusted by 4 and 5 nucleotides. Peaks were called using the Corepeak package of MACS2 (v2.0.10)<sup>240</sup>.

The ATAC profiles of the replicates were merged and context-specific peaks identified using pybedtools<sup>241</sup>. Peaks overlapping the first exon of genes in danRer10 were identified by Pybedtools and used to distinguish between promoter and non-promoter peaks. Non-promoter peaks were associated with the nearest expressed (FPKM  $\geq$  2) genes using bedtools (v.2.25), closestBed function.

In gene desert regions all ATAC peaks up to neighbouring genes must be considered as potentially relating to the central gene despite huge distances involved. In gene dense regions the distance between neighbouring genes is smaller thus the relative expression dynamics of proximal genes is taken into account when associating open chromatin regions with a gene

### 2.5.3. Single Cell RNA Sequencing.

10ss anterior citrine<sup>+</sup> cells were isolated by bisection, cell dissociation and FACS as described previously. FACSARIAIII was set up to aliquot a single citrine<sup>+</sup> cell into each well of a 96 well plate, wells H11 and H12 were intentionally left empty as negative controls. ERCCs are a set of unlabelled polyadenylated RNA controls commonly used in RNA-seq and were developed by the External RNA Controls Consortium (ERCC). ERCCs were included to control for variation in RNA extraction efficiency and loss of material during preparation, and will permit comparison between experiments if repeated at a future date. RNA extraction and library preparation was carried out by R. Williams using the Smart-seq2 protocol as described by Picelli *et al.*<sup>242</sup>. This technique permits high level multiplexing, and restricts sequencing to poly-adenylated RNA. Sequencing was carried out using a NextSeq 500/550 MID output kit (150 cycles) (Illumina) with 75 bp, paired end reads. I carried out indexing to the zv10 genome and mapping of single cell transcription reads using STAR. D. Gavriouchkina carried out quality control and

filtering of data. The SCATER program was used to assess quality of transcriptomic data from each sample through comparison of housekeeping and rRNA gene expression level, plus ERCC level<sup>243,244</sup>. SCATER determined that 11 cell samples should be discarded due to poor quality resulting from lack of depth of sequencing. 2 further samples were eliminated from analysis due to a lack of ERCC spikes. Single cell data has previously been filtered based on cytoplasmic contribution and number of genes associated with mapped reads mapped<sup>245</sup>. Filtering on cytoplasmic contribution would remove any samples that have preferentially amplified rRNAs or mitochondrial RNAs, thus not a true representation of the transcriptome.

**Table 2-1 Results of single cell quality control. Samples to be omitted.**

<b>Well</b>	<b>Reason for removal</b>
H11	Control well with no cell, less than 3000 mapped genes.
A11	Inadequate sequencing depth
A12	Inadequate sequencing depth
B9	Inadequate sequencing depth
B11	Inadequate sequencing depth
C5	Inadequate sequencing depth
H1	Inadequate sequencing depth
G4	Inadequate sequencing depth, less than 3000 mapped genes
H12	Control well with no cell, inadequate sequencing depth, no ERCCs
C12	Inadequate sequencing depth, less than 3000 mapped genes
G1	Inadequate sequencing depth, less than 3000 mapped genes
B12	Inadequate sequencing depth, less than 3000 mapped genes
A4	Less than 3000 mapped genes
D5	Less than 3000 mapped genes
E8	Less than 3000 mapped genes
E12	Less than 3000 mapped genes
F1	Less than 3000 mapped genes
F12	Less than 3000 mapped genes
G8	Less than 3000 mapped genes
G12	Less than 3000 mapped genes
A9	Less than 3000 mapped genes
H8	Less than 3000 mapped genes
F4	Less than 3000 mapped genes, no ERCCs

All of the single cell samples showed >85% of mapped genes related to cytoplasmic RNAs. Islam *et al.* use a 5000 gene cut off<sup>245</sup> for gene number associated with mapped reads, in my single cell experiment only 32 samples show an FPKM greater than 1 for at least 5000 reads, thus it was determined to lower this threshold to 3000

genes. Upon application of this mapping cut-off, 16 samples were identified as failing to map to at least 3000 genes, of these 5 had previously been identified as failing SCATER quality control. Table 2-1 details the omitted samples and the reason for their removal from transcriptomic analysis.

## 2.6. Imaging techniques.

Embryo clutches were screened and imaged on the MVX10 by Olympus. Lightsheet microscopy was carried out on anaesthetized embryos in 1% low melt-point agarose with 4% Tricaine in E3 medium. A Zeiss Z1 Lightsheet was used with single photon excitation and dual signal capture to image embryos.

Confocal imaging was carried out on a Zeiss confocal laser scanning microscope 780 inverted. Multiphoton illumination was used for time-course imaging to achieve greater sample penetration, with non-descanned detectors to improve sensitivity. Z-stack and time course datasets underwent 3d-drift correction processing in Fiji image processing software and were then modelled using Imaris software. Hyperspectral imaging was carried out with single photon excitation and collected emission datasets were processed using the HySP software developed by F. Cultrale. 10-12 embryos were imaged for each hyperspectral experiment.

### 4M Tricaine stock (100mls) pH to 7.0 with 1M HCl.

400g		Tricaine powder
97.9ml		H <sub>2</sub> O
2.1ml	1M	Tris pH 9.0

## 2.7. Primer table.

**Table 2-2 Table of primers**

Primer Id.	Sequence
ptarbac loxp 5'	GCTGTCGGAATGGACGATA
ptarbac loxp 3'	GCAAGTATTGACATGTCGT
Start_of_CitF	gtccggcgaggcgaggcgatgcc
Genomic_F1	GATGTCCTCAGTGGAGTCACCGACATC
Genomic_R2	CATCAGATAATCCCTGGATGATCTTCCTC
Recomb_Stop_F1	caagaggactacaccattccagcctaccgctggacgaaatgccagcggGATGGAAGTGACTACAAGGACGACG
Recomb_Stop_R1	atacctagtgtcagcccagaaaattagtctcagtgctccaatcagCCAGAAGTAGTGAGGAGGCTTT
Recomb_AS_F1	caacttgatttgataaccactttgtacattttctggatcgccggaaggATGGCTGGTGGCCTGAATGACATCTT
Recomb_SA_F1	caacttgatttgataaccactttgtacattttctggatcgccggaaggATGATGGAAAACTGAAATCCGAGCA
Recomb_R1	ctcactcttctgtgccttaattgccatgagtataaagcgcgttacCCAGAAGTAGTGAGGAGGCTTT
Mid_Cit_R	CCCTCGCCGGACACGCTGAACTTGTG
RT-PCR-Fwd	GATTTAGGTGACACTATAGGGTGCAGACCACTGAGCTGTGCAGACCCCC
5' SP6 endogenous scl F	GATTTAGGTGACACTATAGCGACCAGGACGACATGGTCCG
3' T7 endogenous scl R	TAATACGACTCACTATAGGGAGTCCGGCTGACCTCCATACC
5'_SP6_gfi1aa	GATTTAGGTGACACTATAGCGCACAGTTATCATCAGCCC
3'_T7_gfi1aa	TAATACGACTCACTATAGGGCCCCGCTGTGTTCTTTGT
5'_SP6_slc9a3r1	GATTTAGGTGACACTATAGTTTCATCTCCACGGCGAGAA
3'_T7_slc9a3r1	TAATACGACTCACTATAGGGAGCCTGCTGAAGAGACATGT
5'_SP6_cd63	GATTTAGGTGACACTATAGAGGAGGAGCGAAATGTGTCA
3'_T7_cd63	TAATACGACTCACTATAGGGTGACAAAATGCAGGCCAACA
5'_SP6_klf17	GATTTAGGTGACACTATAGGCAGATGCGATGTTGCCTT
3'_T7_klf17	TAATACGACTCACTATAGGGGTGCCTCTTCATGTGCAGAG
5'_SP6_znfl2a	GATTTAGGTGACACTATAGATGAGGGAAACACACACCCGA
3'_T7_znfl2a	TAATACGACTCACTATAGGGGTGTGAGGGAGTGTGCTG
5'_SP6_plk3	GATTTAGGTGACACTATAGTTGTTTACCACAGCTCAGC
3'_T7_plk3	TAATACGACTCACTATAGGGTGGCCTCATTCACTCCATT
5'_SP6_epcam	GATTTAGGTGACACTATAGTGTGGCATTGGTTGATGTGG
3'_T7_epcam	TAATACGACTCACTATAGGGACTGTACTGGCCTTCTGTC
5'_SP6_cfl11	GATTTAGGTGACACTATAGGCCTCAGGTGTAGCGATCA
3'_T7_cfl11	TAATACGACTCACTATAGGGTCCCCCTCAAAGACGTTACA

### 3. Generation and testing of transgenic fish lines

#### 3.1. Generation of BAC transgenic fish lines

##### 3.1.1. BAC selection.

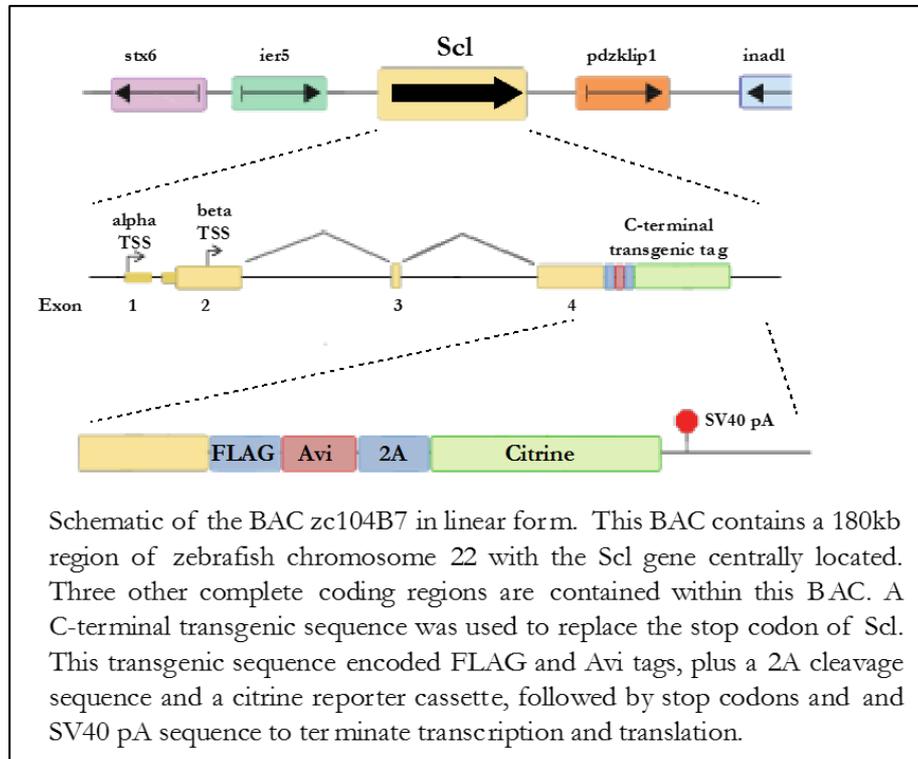
The *scl* transgenic lines generated for this project were designed to enable a range of biochemical investigations, some of which were beyond the scope of this project, but would make the lines useful for future investigations. I aimed to produce zebrafish lines that would express Avi-tagged (biotinylatable) Scl protein, together with a fluorescent reporter, at close-to-physiological levels and in the same cells as the endogenous *scl*.

BACs (bacterial artificial chromosomes) are 50-200kb circular DNA molecules that can contain large genomic DNA regions within the context of a modified backbone, featuring a bacterial origin of replication and partitioning genes to enable stable maintenance. Several zebrafish BAC libraries exist of varying insert sizes (100-350kb), covering the majority of zebrafish genome.

For the purpose of this study I used the zc104B7 BAC that contains a 190kb region of the zebrafish genome, with the *scl* gene located in the centre of this sequence and potentially contains many of the regulatory elements required for endogenous-like expression. Four additional coding sequences are also included in this region (Figure 3-1). Genome alignments across vertebrate species show high conservation of the *scl* loci<sup>155</sup>. In *Takifugu rubripes* the *scl* locus is flanked by unrelated genes to those flanking the locus in the zebrafish and mouse genomes, suggesting that *scl* regulatory regions are contained within the genomic region between the two neighbouring genes<sup>166</sup>. Zebrafish *scl* knockdown can be rescued using a 10.4kb fragment that contains the *scl* locus of *Takifugu rubripes*, which further supports that this region contains all of *scl*'s

regulatory elements. Together this data suggests that this BAC will contain the majority of the regulatory regions that control the expression of the *scf* gene.

**Figure 3-1 Schematic of the *zc104B7* locus and the intended structure of a C-terminally tagged *scf* transgene**



The central location of *scf* within this BAC offers greater protection from the influences of the surrounding chromosomal architecture, once integrated into the genome. As a result variation in expression levels between different experiments should be minimised and more accurately represent endogenous *scf* expression. Three other intact genes are included on this BAC; these are *stx6*, *ier5* and *pdzkclip1*.

### 3.1.2. Construct structure

The *Scf* protein has been previously tagged at both N- and C- terminal ends, in efforts to investigate its mode of action<sup>246,247</sup>. The complete *Scf* protein structure has yet to be solved, thus it is difficult to predict the implications of adding extra domains, however small. bHLH DNA binding domain is encoded by the sequence within the exon 4 and situated 183 amino acids from N- and 87 amino acids from C-

terminus of the full length protein. A study by Cai *et al.* indicated that the presence of C-terminal tags did not prevent protein complex formation required for mediating repression via interactions with Eto2<sup>246</sup>. C-terminal tagging of Scl has also been used to perform ChIP experiments in cell culture, indicating that the tags did not impair association of Scl protein with DNA and remained available for immunoprecipitation following the crosslinking of the cellular material<sup>247</sup>. Similarly, N-terminally-tagged Scl protein has been used to investigate this factor's response to Notch signalling in culture<sup>175</sup>. There is currently no evidence in the literature suggesting that the inclusion of tag sequences at either end of the Scl protein would affect its functional activity. I generated two constructs tagging zebrafish Scl protein, one at the N- and the other at the C-terminus.

The tag sequence combined several motifs - Avi tag, TeV site, FLAG tag and a linker sequence. In both the N- and the C-terminally tagged transgenes the linker sequence is directly adjacent to the *scl* coding sequence such that the unstructured linker peptide it encodes separates the potential functional domains within the Scl protein from the Avi-tag placed in the most distal position.

When expressed in the same cells as the specific bacterial biotin ligase BirA, a single lysine residue within the Avi tag is efficiently biotinylated, yielding a tagged Scl protein that can be used in biochemical investigation of Scl molecular and cellular function, due to its high affinity association with streptavidin. Biotin-streptavidin interactions are one of the strongest non-covalent interactions in nature ( $K_d \sim 10^{-15}$ ), permitting streptavidin-based affinity purification of Scl protein targets and their interacting entities (nucleic acids, proteins etc.) with exquisite stringency.

The TeV site within the tag sequence encodes a short peptide specifically cleaved by the Tobacco Etch Virus protease. This sequence is included so that biotinylated-Scl can be digested off streptavidin beads. This specific elution following transcription

factor pull-down permits investigation of the protein binding partners or post-translational modification, rapidly from an *in vivo* source and eliminates background in mass spectrometry experiments, coming from endogenously biotinylated proteins and the streptavidin itself. The FLAG tag allows confirmation of expression of exogenously tagged Scl protein, its subcellular localisation and pull-down (much less effectively than by biotin-ChIP), in the absence of biotinylation. Finally a linker sequence of unstructured peptide was included to minimise the effect of the presence of the tag on the function of Scl, and maximise its availability for pull-down.

After the *scl* coding sequences a 2A sequence encoding *Thosea asigna* virus peptide<sup>248</sup> was included followed by the YFP-variant, citrine, fluorescent reporter gene sequence. Transcription from this transgenic *scl* loci produces a single transcript containing tagged *scl* sequence, 2A sequences and then the citrine reporter sequence. As this transcript is translated the 2A peptide allows either a self-cleaving<sup>249</sup> or ribosome-skipping mechanism<sup>250</sup> to occur, causing physical separation of Scl protein from the fluorescent reporter protein.

### 3.1.3. BAC transgenesis.

To generate *scl*-tagged BACs, the donor cassette containing the tag sequence and reporter, followed by FRT site flanked kanamycin selection gene was recombined into the selected BAC backbone using lambda prophage homologous recombination system available in the SW105 bacterial background according to the previously published protocol<sup>251</sup>. I then used Tol2-mediated transgenesis to integrate recombiner BACs into zebrafish genome<sup>252,253</sup> through co-injection of the BAC with *tol2 transposase* mRNA into single celled zebrafish embryos. The Tol2 system is an autonomous transposon originating from the medaka fish (*Orizyas latipes*) and has

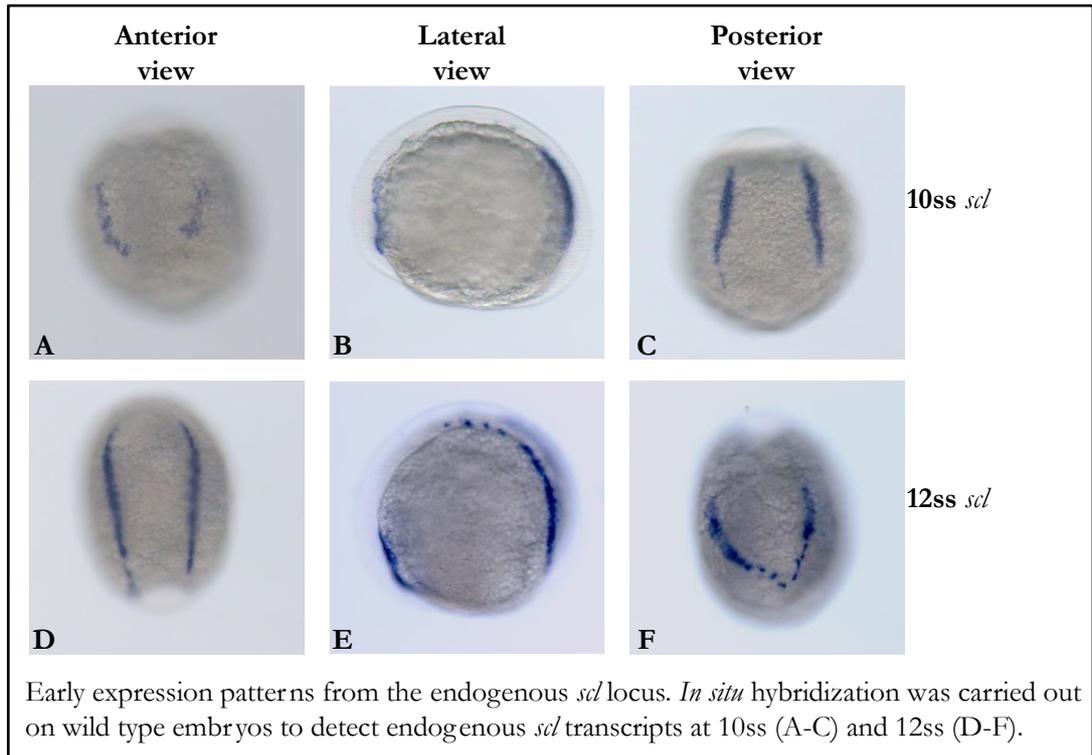
been shown to be active in all tested vertebrates<sup>254-256</sup>. The 190kb tagged *scf* genomic loci, contained within the BAC was flanked by terminal inverted repeats (TIR), which have previously been identified as necessary and sufficient for transposition<sup>257</sup>. When the Tol2 transposase protein is produced in cells, Tol2 transposase recognises the TIR motifs and mediates a single integration of the sequence they flank into the genome, with very low sequence preference<sup>258-260</sup>.

To pass the transgene onto the next generation the BAC must be integrated into the genomic DNA of a germ cell. Germ cells are segregated early in embryogenesis, thus efficient transgenesis requires injection within 5 minutes of zebrafish spawning. Injection at the single cell stage maximises the chance that all future cells receive the BAC plus mRNA or protein for Tol2 transposase. Due to differences in segregation of these factors between daughter cells, mosaicity of integration is seen in the F<sub>0</sub> generation. Three different tagged *scf* BACs were injected, each displaying similar phenotypes and transient expression patterns.

Approximately 60% of injected embryos died within the first 24 hours (see Figure 3-3 section A), this is not seen upon BAC injection alone, suggesting that Tol2-mediated integration can cause fatal genome disruption. This initial death rate was later used to monitor the quality of *tol2* mRNA. Of the surviving embryos a further 75-79% died before reaching 5 days post fertilisation, most likely due to gross morphological abnormalities consistent with an *scf* overexpression phenotype. The *scf* overexpression phenotype documented by Gering *et al.*<sup>17</sup> includes increased number of *scf*<sup>+</sup> cells in the ICM, cardiac oedema plus dorso-ventral axis deformations. Surviving embryos were entered into the system and grown to maturity- these embryos showed expression of the citrine reporter yet did not display any symptoms of the aforementioned *scf* overexpression phenotype. Post 5dpf survival to adulthood was 80- 90%. Following *scf*BAC:*scf-flag-tev-avi-2A-citrine-sv40pA* and *tol2* mRNA

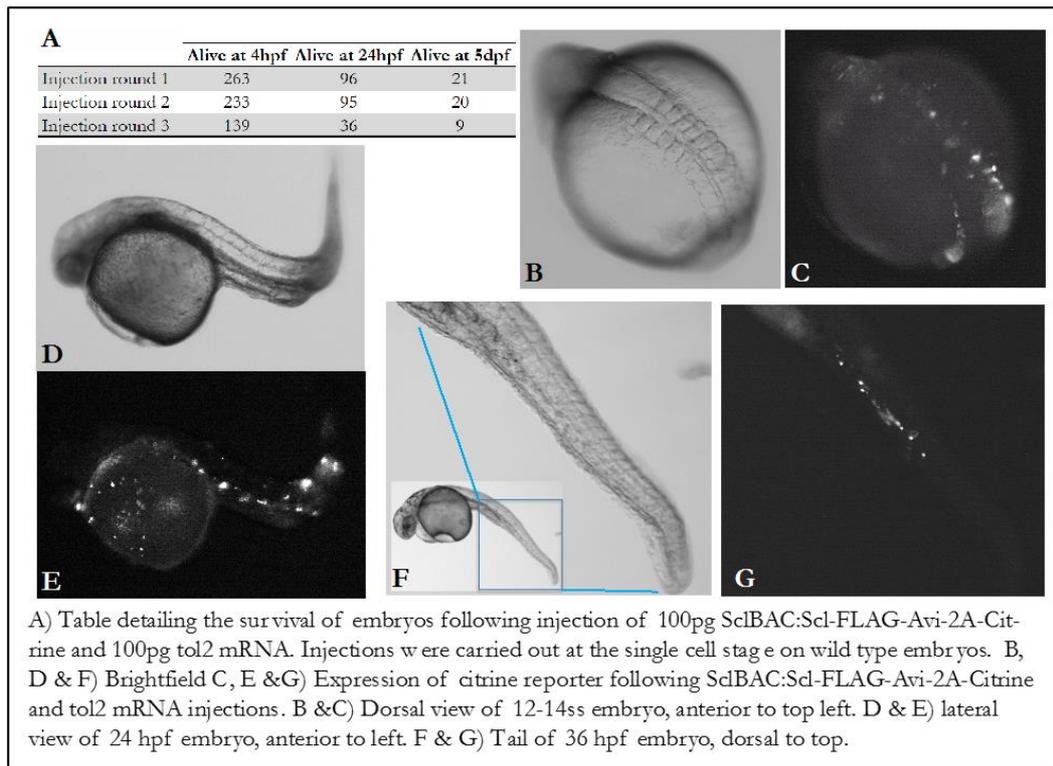
injections, transient expression pattern of the citrine reporter from the *scl*BAC:*scl*-*flag-tev-avi-2A-citrine-sv40pA* integrated into somatic cells, was consistent with endogenous *scl* expression, as pairs of bilateral stripes in the ALM and PLM, as previously reported<sup>17</sup> (See Figure 3-2 and Figure 3-3).

**Figure 3-2 Endogenous *scl* expression patterns as determined by *in situ* hybridization.**



Ectopic expression was also visible in this F<sub>0</sub> generation, this however has previously been reported following BAC transgenesis in zebrafish, results from transcription from the circular, non integrated BACs and is lost upon transmission to the F<sub>1</sub> generation (A. Kenyon- unpublished data).

Figure 3-3 Initial results from injection of *scIBAC:scl-flag-tev-avi-2A-citrine-sv40pA* plus *tol2* mRNA into F<sub>0</sub> generation embryos.



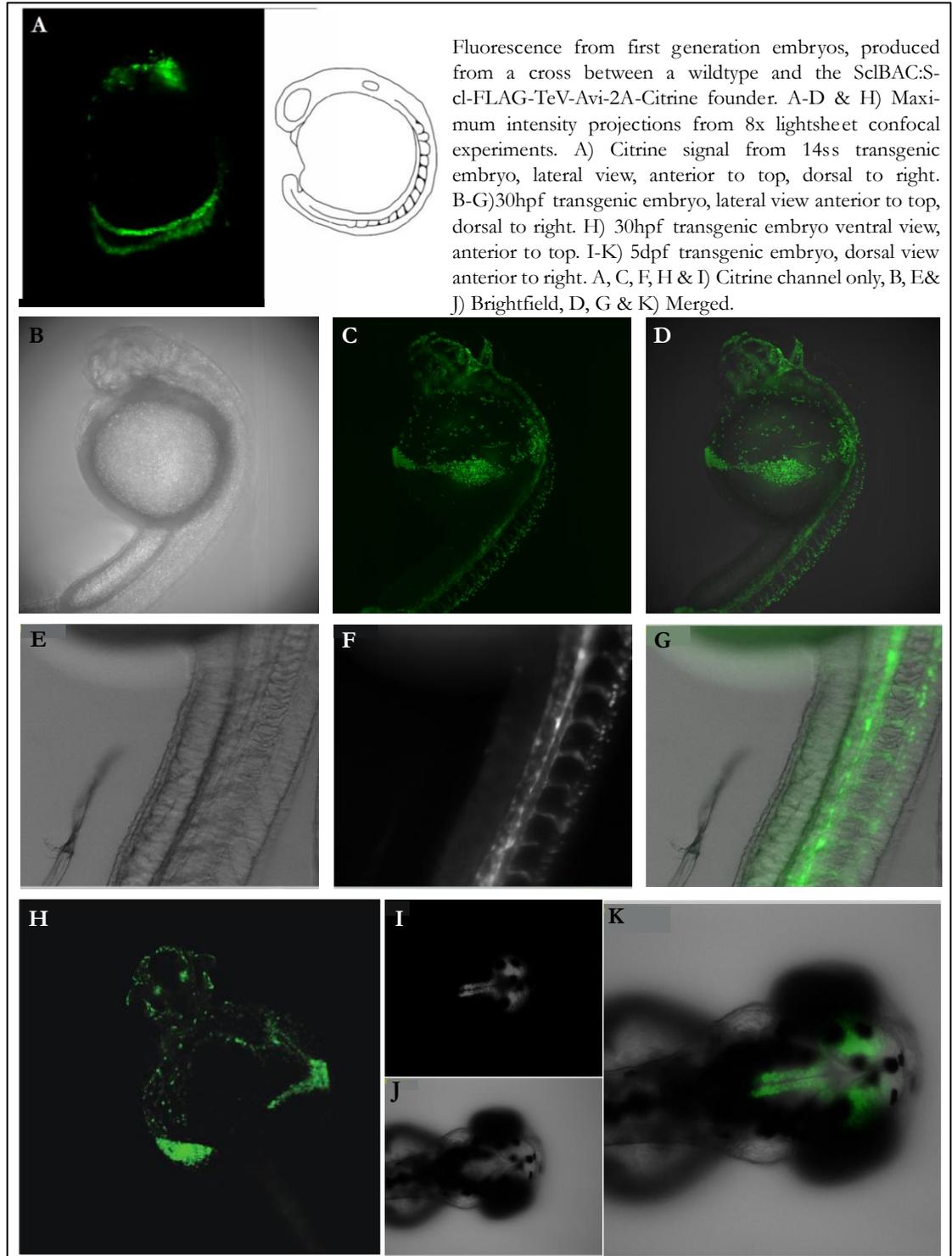
#### 3.1.4. Screening for germline transmission.

Transmission to the F<sub>1</sub> generation is essential for line generation. Screening of the F<sub>0</sub> generation for germline transmission was carried out primarily by genomic PCR screening. This molecular technique has the advantage of detecting the presence of the transgene in the genome, regardless of the expression level or the number of offspring carrying the transgene. Multiple primer pairs were tested and those that generated the least non-specific products yet still amplified the positive control were used (See Figure 2-1).

Following identification of putative founders by genomic screening, subsequent clutches from these fish were examined under fluorescent microscope to identify those F<sub>1</sub> offspring that have inherited the BAC in their genome and were thus expressing the fluorescent reporter. Embryos were dechorionated and assessed for citrine fluorescence at the 24hpf stage, when expression is seen to be maximal by *in*

*situ* hybridisation studies. The observed *scf* driven citrine fluorescence is shown in figure 3-4.

**Figure 3-4** Fluorescent screening for *scfBAC:scf-flag-tev-avi-2A-citrine-sv40pA* transgenic F<sub>0</sub>s with germline transmission of transgene.



Expression matched previously seen endogenous *scf* expression patterns (See Figure 3-2 and Figure 3-3) and showed greater intensity than that shown by the F<sub>0</sub> transients. Ectopic expression previously seen in the F<sub>0</sub> generation was no longer present. Positive embryos were recorded and grown to 5dpf, at which point they were entered into the aquatics system and grown to maturity. No obvious mutant or *scf* overexpression phenotype was visible and embryos showed 100% survival to maturity (60 dpf).

Outcrosses of the F<sub>1</sub> generation gave proper Mendelian ratios (50% citrine<sup>+</sup> embryos when assessed between 16-22ss), suggesting a single site of BAC integration in the founder.

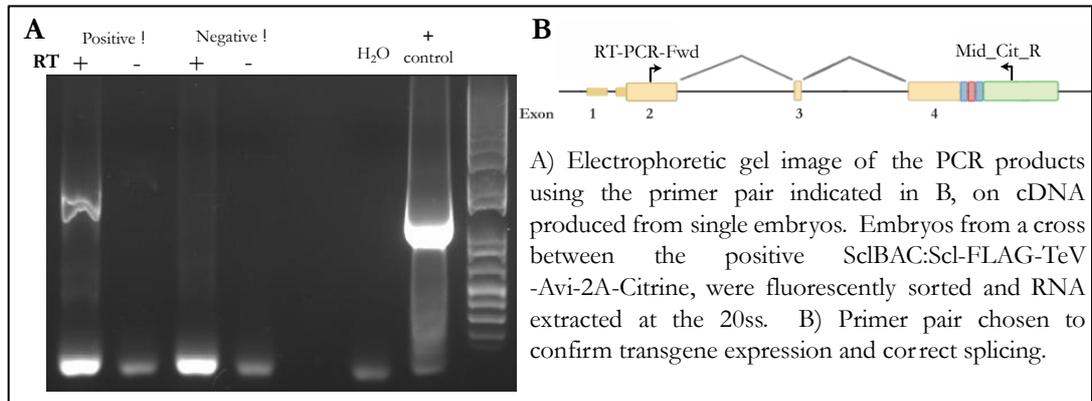
The Tol2 mechanism of transposition is based on recognition of a target site sequence ((C/G)TTATAA(G/C)), which occurs roughly 350,000 times in the zebrafish genome. Each transposition event occurs independently and results in duplication of the target sequence following successful transposition<sup>258</sup>. Multiple tol2-mediated BAC integrations within a single transgenic line have not been observed to date (unpublished data). Despite the target sequence being retained its frequency within the genome makes multiple BAC insertions at the same site highly improbable.

#### 3.1.5. Confirmation of transgene expression *in vivo*.

RNA was extracted from 50 citrine<sup>+</sup> 20ss *scf*BAC:*scf*-FLAG-TeV-Avi-2A-Citrine embryos and used to produce cDNA by reverse transcription. Reverse transcription-PCR was then carried out to probe for the presence of the transgenic transcripts within the total RNA of these cells. Figure 3-5 shows the results of this analysis, with amplification of a product matching in size to the positive control. This analysis confirms expression of the *scf* coding regions within the same transcript

as the citrine reporter sequence, in RNA isolated from transgenic embryos. Figure 3-5 also demonstrates that this transgenic transcript undergoes the same splicing events as the endogenous *scf* transcripts, to yield a final mature transcript.

**Figure 3-5 Electrophoretic gel image of RT-PCR products, probing for transgenic transcripts.**



These screening approaches confirm that the F<sub>1</sub> generation carry a copy of the *scf* transgene at the level of DNA, RNA and protein (as concluded from citrine protein fluorescence). This transgene exists in addition to the endogenous *scf* gene and its products, thus represents an overexpression when compared to wild-type embryos.

## 3.2. Characterization of transgenic fish line.

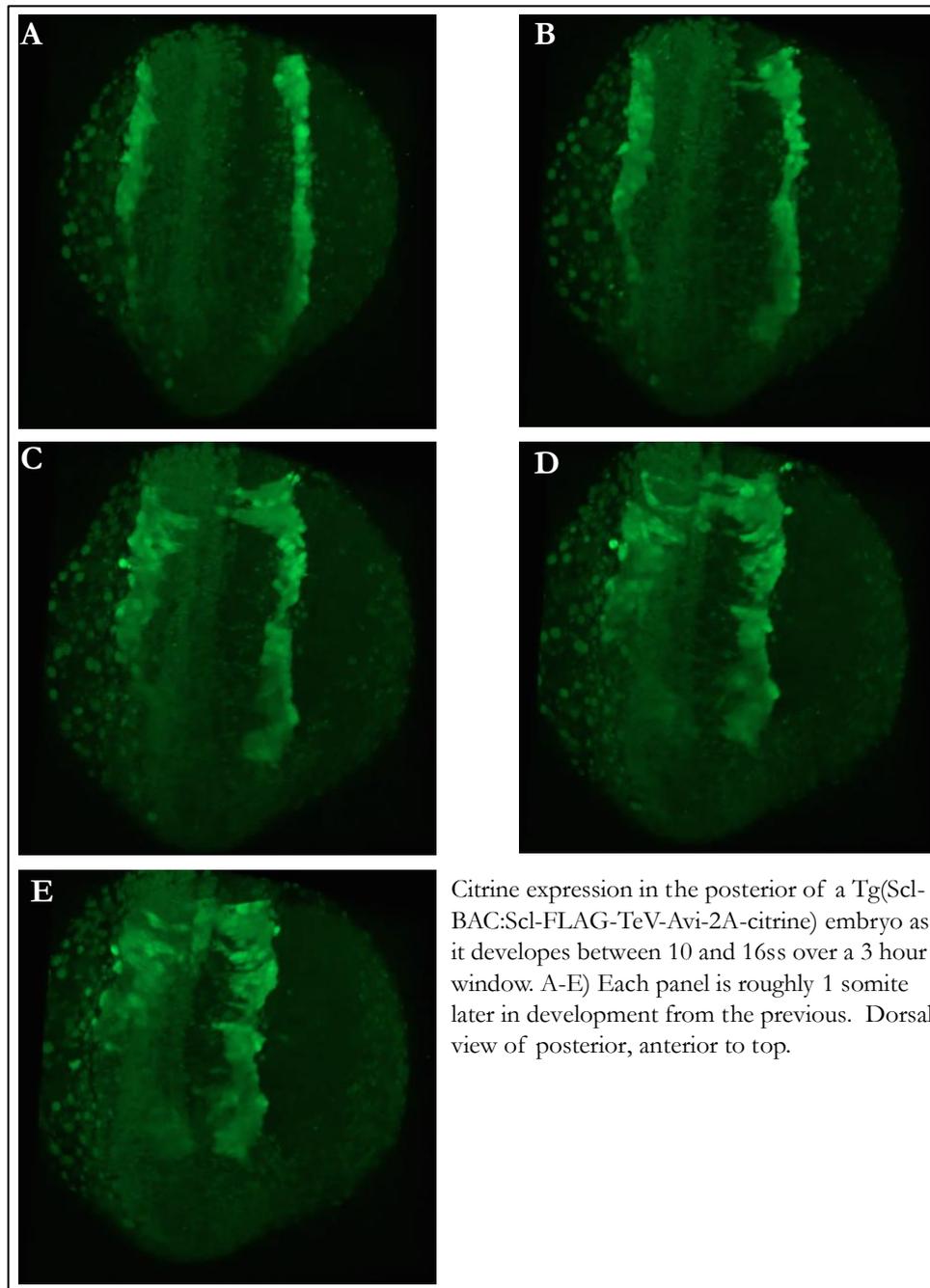
### 3.2.1. Characterisation of transgene expression *in vivo*.

Citrine expression was readily observable and showed great similarity with previously published reports and my own *in situ* experiments, of *scf* expression patterns<sup>17</sup>. Endogenous *scf* transcript can be detected as early as the 2ss, however a delay of ~1hr for translation and protein folding can be expected, this correlates to an increase in 3ss/hr.

The posterior citrine<sup>+</sup> positive populations extended towards the anterior and expression of citrine was observed in the anterior of the embryo as two stripes. These pairs of *scf* populations were imaged over development to follow the citrine

expression pattern and correlate this with known *scl* expression patterns and functions.

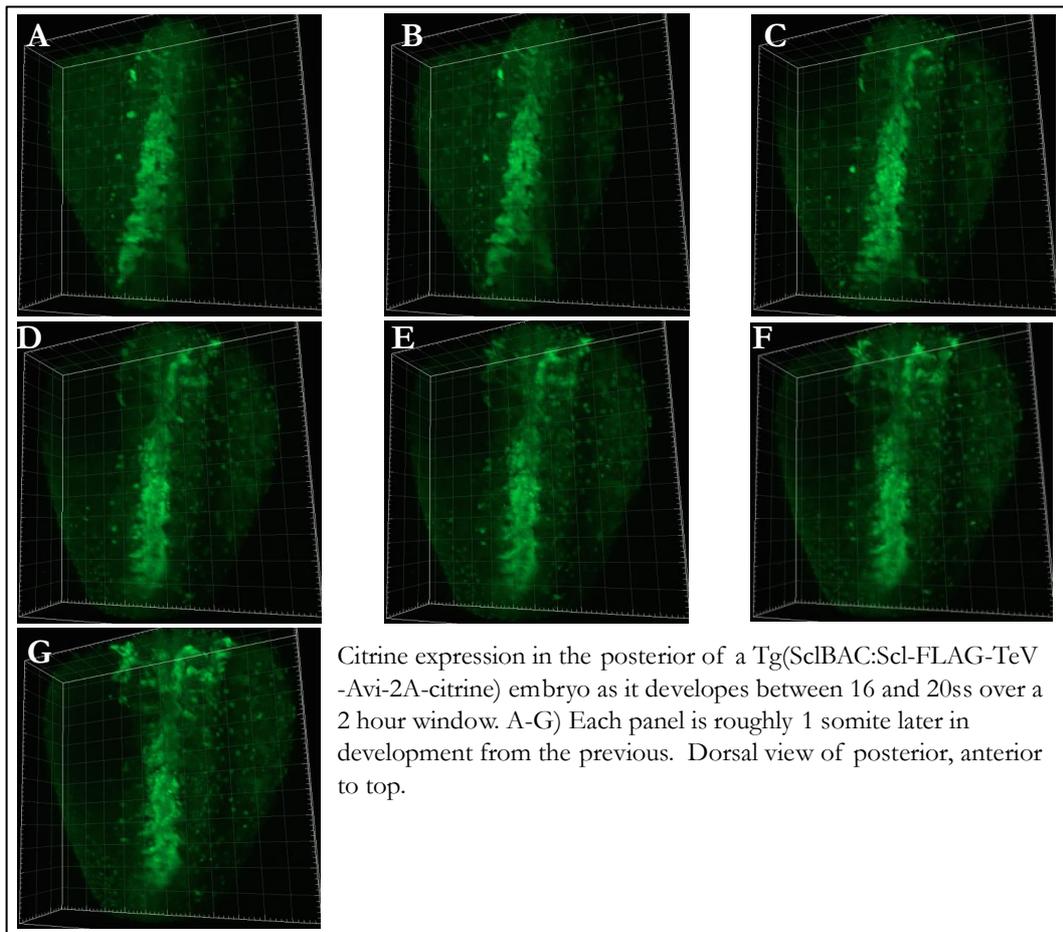
Figure 3-6 Time series showing *citrine* expression in the posterior of the 10-16ss *sclBAC:scl-flag-tev-avi-2A-citrine-sv40pA* transgenic F<sub>1</sub> embryo



Comparison between figure 3-6, figure 3-7 and figure 3-2 show that the citrine reporter closely copies the endogenous expression pattern in the posterior of the embryo during early haematopoiesis.

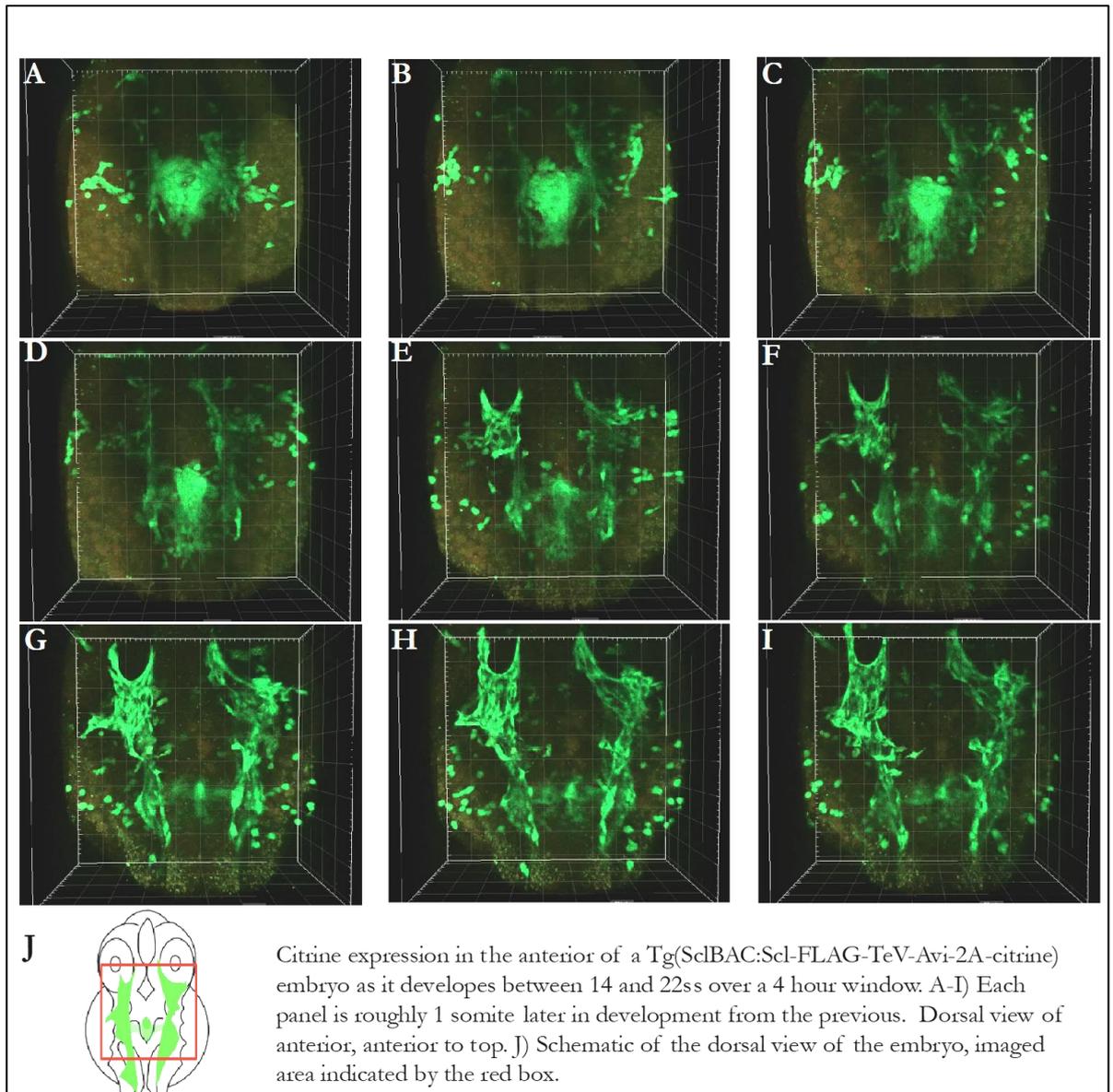
Figure 3-7 shows how as development progresses the posterior bilateral stripes migrate to the midline, coalesce and form the intermediate cell mass (ICM). This has previously been published by Zhang and Rodaway<sup>198</sup>, further supporting that use of the *scI* BAC drives transgene expression in a fashion that closely mimics endogenous *scI* expression.

**Figure 3-7** Time series showing *citrine* expression in the posterior of a 16-20ss *scI*BAC:*scI*-*flag*-*tev*-*avi*-2A-*cit*rine-*sv40pA* F<sub>1</sub> embryo.



Anterior expression of *scI* was identified and followed through the development of the heart field and the initial stages of cranial vascularisation. Figure 3-8 shows the movement of citrine positive cells out of the lateral plate mesoderm, into the heart field and over the yolk sac. Strings of citrine<sup>+</sup> cells extend into the far anterior of the embryo forming the initial blood vessels around the developing eyes.

Figure 3-8 Time series showing *scl* driven *citrine* expression in the anterior of a 14-22ss *sclBAC:scl-flag-tev-avi-2A-citrine-sv40pA* transgenic F<sub>1</sub> embryo.



*Scl* contributes to neural development and as evident by the restriction of *scl:citrine* expression to the hindbrain by day 5 (panels I-K figure 3-4). Neural expression of *citrine* is first observed at 22hpf in the dorsal root ganglion and increases in intensity over days 2-4 of development while haematopoietic and vascular expression decreases. No neural expression of *citrine* is observed in either the anterior or posterior before the 22ss as evident in figure 3-6 and figure 3-8.

Characterisation of this transgenic line suggests that despite the presence of a third allele of *scl* in the form of the transgene, citrine expressing cells arise and behave in a

fashion previously reported for endogenous *scf* expressing populations. No cardiac oedema, axis deformation or expansion of the ICM was observed in any of the F<sub>1</sub> generation embryos or F<sub>2</sub> generation embryos produced from F<sub>1</sub> in-crosses (data not shown). Together this suggests that the over abundance of *scf* coding sequences, is moderated by down regulation at the transcriptional or protein degradation level, or that the increase in Scl protein activity is tolerated by the regulatory networks active within these studied populations. As previously discussed, transcription from the *scl* locus is known to be subject to an auto-regulatory loop<sup>139,162-165</sup> (see section 1.2.5). Such a regulatory kernel may result in decreased transcription occurring at each of the alleles in the transgenic, compared to wild-type, yet maintain total *scf* transcript levels to match the endogenous situation.

### 3.3. Preliminary biological sample collection and optimisation

#### 3.3.1. Choice of time points

Here I propose to investigate regulatory landscape of *scf*-expressing cells in 2 different spatial developmental contexts (anterior and posterior, corresponding roughly to myeloid and erythroid programmes, respectively) and at two different time points (early progenitor and later differentiation stage). Ideally this project would investigate the *scf* activity of the first cells expressing this master regulator, however these are very few in number and technically difficult to collect. Currently collection of a sufficient number of *scf*<sup>+</sup> cells at the 3ss would be unfeasible for biotin-ChIP and ATAC investigation. The added difficulty resides with the fact that fluorescent reporters, produced in *scf*-positive cells, that were used for cell isolation have a distinct maturation/folding time before they can be detected by FACS. I also used FACS analysis to assess the number of *scf*<sup>+</sup> cells present and collectable by flow cytometry at early time points, figure 3-9. I chose the 10-somite stage (ss) as my early

time point, as the greater cell number would make sample collection swifter and thus more developmentally accurate, even if the population were not the earliest *scf* population to arise.

**Figure 3-9 Preliminary FACS isolation of *citrine*<sup>+</sup> populations**

<b>A Intact and entire embryos</b>				
Developmental stage of sample	8ss	10ss- 1	10ss- 2	10ss- 3
Number of embryos dissociated	68	196	40	174
Number of citrine positive cells	10793	85407	10213	60074
% citrine positive cells out of total cell number	0.5	0.7	0.6	0.7
% survival	23.5	48.9	24.1	36.7
Total cell number	45928	174656	42378	163689
Number of citrine positive cells/ embryo	675	891	1059	941
Number of citrine positive cells isolated/ embryo	159	436	255	345

<b>B Anterior cells</b>				
Sample	1*	2	3	4
Number of embryos dissociated	201	67	113	100
Number of citrine positive cells	10504	1569	7743	6607
% citrine positive cells out of total cell number	0.2	0.4	0.4	0.4
% survival	44	63	45.2	51.4
Total cell number	23873	2490	17131	12854
Number of citrine positive cells/ half embryo	119	37	152	129
Number of citrine positive cells isolated/ half embryo	52	23	69	66

<b>C Posterior cells</b>				
Sample	1*	2	3	4
Number of embryos dissociated	~100	67	113	100
Number of citrine positive cells	15063	6233	22480	7521
% citrine positive cells out of total cell number	0.9	0.9	1	0.4
% survival	38.2	55	48.3	39.8
Total cell number	39432	11333	46542	18897
Number of citrine positive cells/ half embryo	~400	169	412	189
Number of citrine positive cells isolated/ half embryo	~150	93	199	75

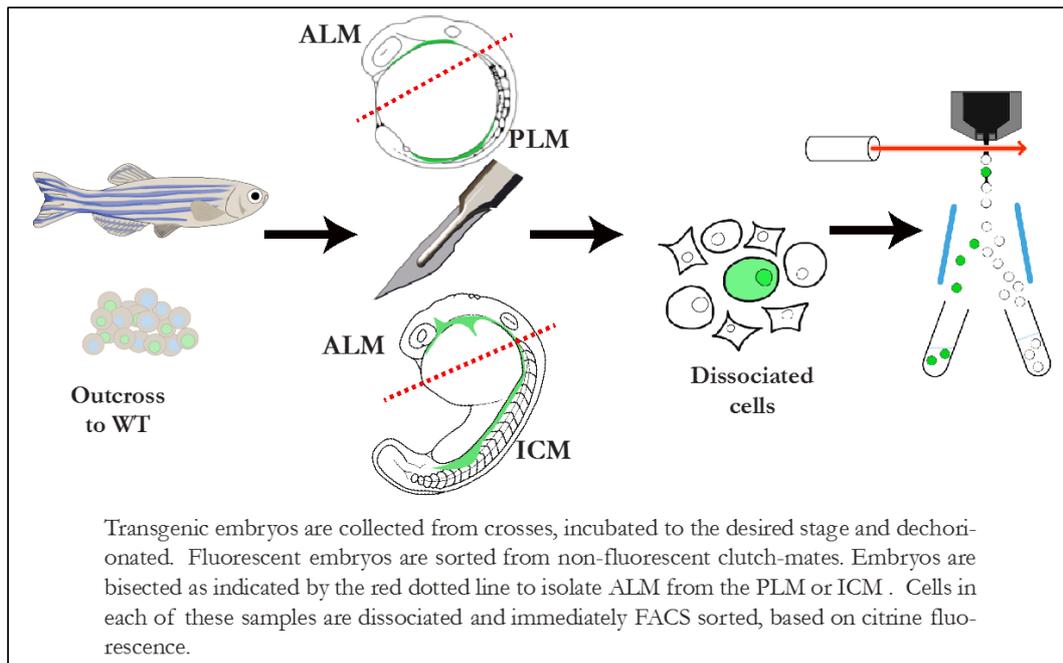
Details of the citrine<sup>+</sup> population size and efficiency of isolation by FACS during early haemangioblast development. A) Cell numbers for 8 and 10ss whole embryo (i.e. without bisection) samples. B) Cell numbers for 10ss anterior samples following embryo bisection at the 8-9ss. C) Cell numbers for 10ss posterior samples following embryo bisection at the 8-9ss. Sample 1\* included in tables B & C was produced by bisection and dissociation of 201 embryos, 201 anteriors or 201 posteriors. During the posterior run the system encountered an error and the sort had to be halted roughly halfway through.

Circulation in the zebrafish commences between 24 and 30hpf, thus it is necessary for the later stage investigation into the myeloid and erythroid networks to occur before these populations become mixed. As development progresses the *scf*<sup>+</sup> population also becomes more complex and diverse as these cells develop down different haematopoietic lineages. The results of any population study such as an anterior transcriptome would therefore be a result of a highly diverse population and thus less informative for regulatory network analysis.

The role of *scl* in neural development would add further complexity and noise to any study into the early haemangiogenic roles of *scl*, if these neural cells cannot be removed from the studied population. 22ss is the onset of *citrine* expression in the dorsal root ganglion, the first neural cells observed to express *scl*. Thus, by choosing 20ss as my later time point of analysis I removed any neural component of *scl*'s activity from my analysis of this factor's contribution to a haemangiogenic program. In addition this stage of development avoids combining anterior and posterior originating *scl*<sup>+</sup> populations due to the initiation of circulation. At 20ss the *scl*<sup>+</sup> population has dramatically increased in cell number (figure 3-9) and the cell morphology of transgene expressing cells suggests that these cells have further developed down vascular, myeloid or erythroid lineages since the 10ss. Thus the 20ss populations should be suitable representations of later stages of primitive haemangiogenesis while minimising the complications of cellular diversity.

Embryos will be bisected to separate the early haematopoietic *scl*<sup>+</sup> populations of the anterior and the posterior at both the 10 and the 20ss. FACS permits the isolation of *citrine*<sup>+</sup> cells from these samples in numbers sufficient to perform genome wide profiling of transcripts and areas of open chromatin (see figure 3-10). This rapid isolation process combined with protocols adapted for small cell numbers allows the investigation of dynamic haematopoietic and vascular activity, directly from an *in vivo* source. Physical isolation of the *scl*<sup>+</sup> populations has enabled me to describe these developmentally important populations and compare myeloid progenitor profiles to those of erythroid progenitors.

Figure 3-10 Schematic detailing the collection of spatially separated *scl*-expressing populations from zebrafish embryos.



#### 3.4. Chapter 3 summary.

Towards my aim of investigating the role of early haematopoietic and vascular progenitors marked by the expression of *scl*, I have generated a zebrafish line that expresses transgenic *scl* tagged with *avi* and *flag* tags and a *citrine* reporter gene.

To test the effect of the presence of a tag on the termini of the Scl protein, I cloned N- and C –terminally tagged *scl* constructs and assessed their transient expression and activity *in vivo*. I assessed expression systems for their ability to accurately mimic endogenous *scl* expression and I chose to use a bacterial artificial chromosome containing 190kb of the zebrafish *scl* loci, to drive transgenic *scl* expression for this project. I used Tol2 mediated integration to successfully introduce the transgenic *scl* loci into the zebrafish genome. Zebrafish were identified that carried the transgene in their germ cells and positive F<sub>1</sub> generation embryos were used to characterise and form the transgenic line. Expression of a citrine reporter under the control of *scl* regulatory regions was observed to copy the expression patterns of endogenous *scl*.

This tight overlap of expression enabled the terms *citrine*<sup>+</sup> and *scl*<sup>+</sup> cells to be used interchangeably, when referring to this transgenic line. I used live embryo imaging to follow the developing *scl* populations from the 4ss to 22ss, and observed significant migration of the population in the anterior of the embryo. FACS analysis was employed to assess *citrine*<sup>+</sup> cell number per embryo and also for the anterior and the posterior of the embryo separately. In review of my live embryo imaging, FACS analysis and published literature, I chose to collect embryonic samples at the 10 and 20ss. These time points permitted the comparison of the initiation of primitive haematopoiesis and the development of distinct lineages, from an *in vivo* setting. Through bisection of the embryos, a comparison of the anterior versus the posterior *scl* expressing populations was also possible. This spatial comparison allowed the diversity of the two *scl* populations to be investigated and for the description of myelopoietic and erythropoietic cell profiles separately, due to their distinct spatial origins in the zebrafish embryo.

## 4. Investigating early transcriptional variation within *in vivo* $scf^+$ populations.

### 4.1. Introduction

Spatially distinct  $scf^+$  populations develop into blood and vasculature as embryogenesis progresses, each contributing to a different extent to a variety of lineages. Although ALM and PLM  $scf^+$  populations share the expression of  $scf$ , the question is raised whether the  $scf$ -transcriptional profile varies with cellular context. This is particularly pertinent question given that Scl, similar to many key developmental factors, displays a range of different biological roles.

I have developed approaches to purify and collect samples of  $scf^+$  cells for a range of embryonic contexts, on a scale required for genome-wide analysis. Using RNA sequencing of early cell populations I have addressed how they differ and whether at 10ss spatially distinct  $scf$  expressing populations are still at the common progenitor stage or have already begun to diverge. I investigated the difference between early  $scf^+$  contexts of interest using differential expression analysis of poly-A selected RNA for each context. Differential expression analysis was used to describe the factors expressed in each of the early  $scf^+$  populations studied here and compare their transcriptomes to investigate spatial and temporal differences.

To focus on  $scf$  cell specific transcriptomes the sample collection process was repeated, except that this time citrine negative (therefore transgenic  $scf^-$ ) cells were collected from each of the studied contexts. These citrine negative cell samples were then processed for RNA-seq, to provide control background transcriptomes for each of the studied  $scf$  contexts. I hypothesise that these anterior or posterior, 10 or 20ss  $scf^-$  cells have similar housekeeping and general developmental gene expression as the  $scf^+$  cells from the same cellular context. Thus through comparison of transcriptomes for citrine positive and citrine negative cells from the same

embryological origin, cell type specific differences between these key early contexts will be more identifiable.

#### 4.1.1. Overall expression levels of *sc<sup>+</sup>* contexts.

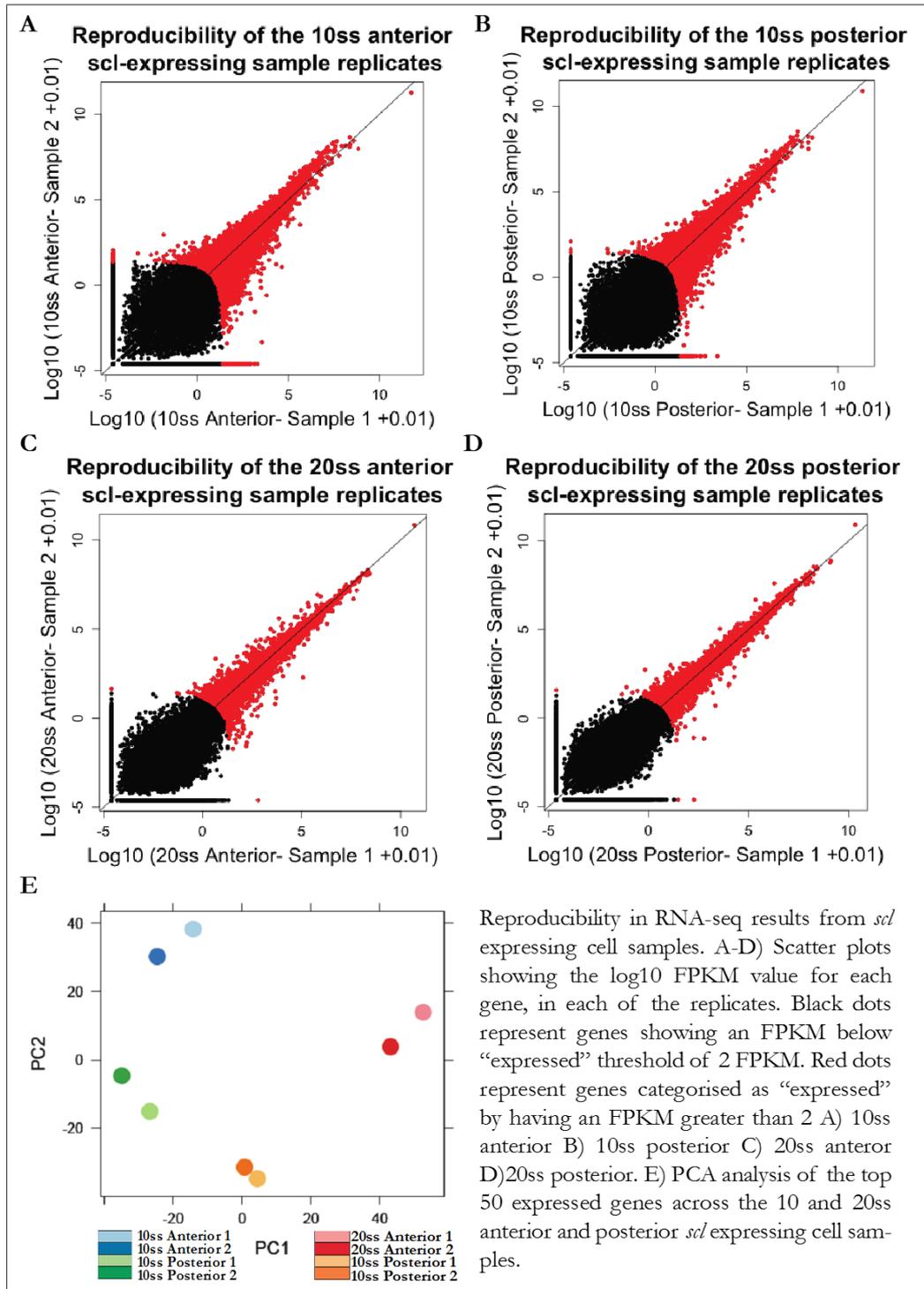
In these datasets a wide range of expression levels between different genes are visible; this variation arises from the detected level of transcription from each gene combined with the percentage of each population that is expressing the gene. As a result in population data such as this a low level of sequenced transcripts for a given gene could either be a result of widespread low expression, or high expression in a limited number of cells.

Transcripts of greater length will have more reads mapping to them than shorter transcripts if present in the same molar concentration. Fragments per kilobase mapped (FPKM) was used as the metric of expression level for all analyses performed, this measure accounts for the length of transcripts, allowing meaningful comparison of expression level between genes of different size.

I have collected and analysed by RNA-seq a minimum of 2 samples for each developmental context analysed (10ss anterior, 10ss posterior, 20ss anterior, 20ss posterior). Biological variation was minimal between replicates, despite batch and sample size differences as shown in figure 4-1. PCA analysis of the top 500 expressed genes from my RNA-seq results suggests that all four contexts have distinct profiles, even at the 10ss.

In my analysis of each early *sc<sup>+</sup>* population genes displaying FPKM values greater than 2 were considered as expressed. This has previously been used in similar analyses as an expression cut-off value<sup>261</sup>.

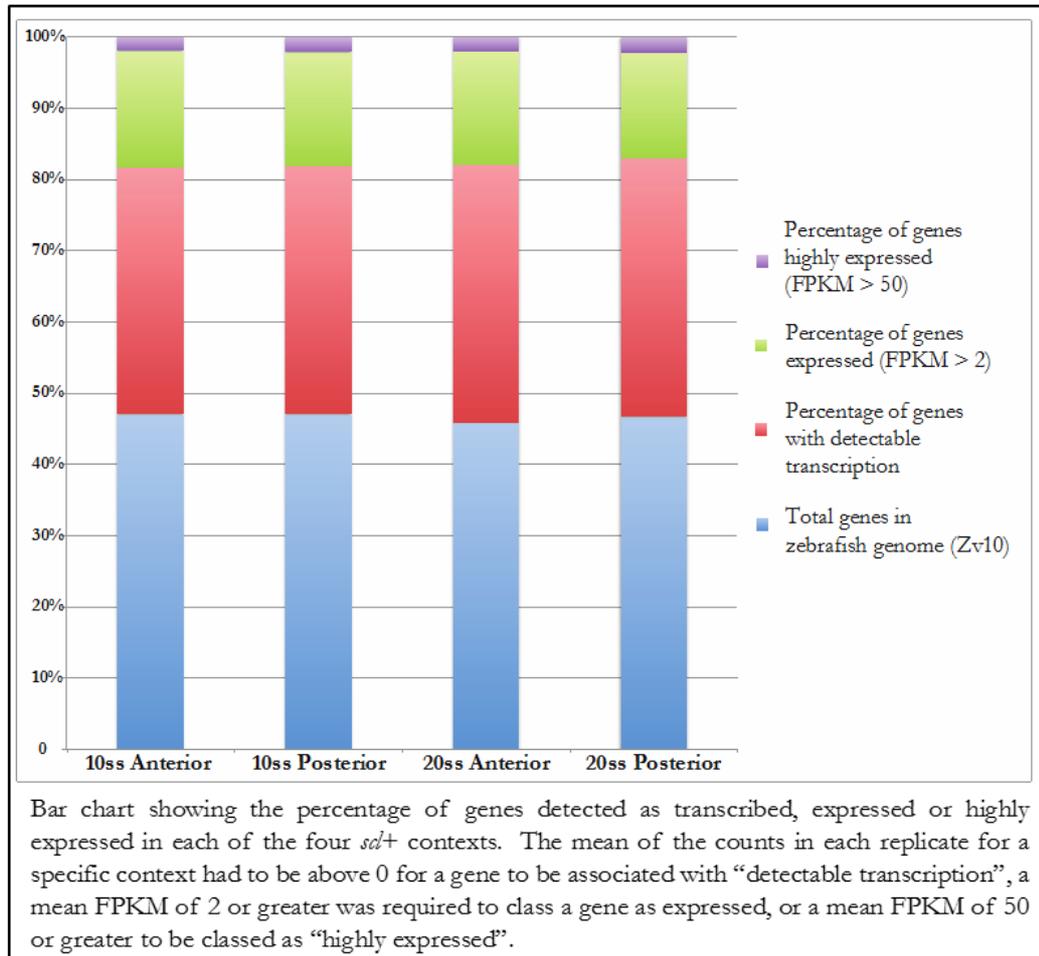
Figure 4-1 Scatter plots of replicates and PCA analysis of early *scl* expressing cells.



50 FPKM was chosen as a cut-off for highly expressed genes as in each sample this represented approximately the top 10% of expressed genes. An expression level cut-off was used to distinguish between highly expressed genes and lowly expressed

genes, to identify which level of gene expression contributed most significantly to specific cellular identity.

**Figure 4-2** Bar chart showing the overall number of genes expressed in each *scf*<sup>+</sup> context.



All four *scf*<sup>+</sup> contexts show similar numbers of genes with detectable transcripts, expressed genes and highly expressed genes (figure 4-2). This suggests that all four contexts are similarly biologically active and diverse. Although the 10ss contexts are the progenitors of their 20ss counterparts, the size of the transcriptomes are comparable suggesting that developmental differences arise from a different selection of genes being expressed rather than the additional onset of transcription at a significant number of genes.

#### 4.1.2. Uses and limitations of gene ontology analysis

Gene ontology (GO) is the association of a given gene with a specific biological or molecular function. This can be used to assess potential biological or molecular trends when investigating a large gene sets, such as those identified as over or under expressed in certain conditions. The accuracy of this preliminary investigation is entirely reliant on the accuracy of the gene annotation database. One such database, ZFIN, contains gene information including ontology relating to known zebrafish genes. I have used the analysis tool, PANTHER (protein annotation through evolutionary relationship), to analyse this database to identify significant enrichment of biological terms within a provided gene list <sup>262</sup>.

The PANTHER tool identifies over or under representation of certain protein classes by comparing the percentage of genes associated with each protein class for the entire genome and the percentage produced from the provided gene set. This tool also provides a p-value of significance for the over or under representation of a particular gene class occurring randomly.

It is important to note that the GO classification system has several disadvantages: first it is based on a database of current published data and knowledge of protein classes and biological functions. As a result differences in nomenclature do not necessarily relate to differences in biological function. Second, a single factor can be annotated with multiple classes or functions, given that proteins can have multiple functions and contribute to a variety of biological processes in nature. However upon classification, each of these GO terms would be presented as equally valid, though potentially erroneously from differences between the context of interest and the context that was studied for the annotation. Additionally GO terms do not distinguish between direct and related functions. A common result of this is that a single factor may be annotated as contributing to processes A, B, C & D- this factor

may act distinctly in each of these processes or be a key factor to process A, which is required for processes B, C & D. It is especially important to be wary of drawing conclusions from genes associated with key developmental decisions, as these may be indirectly necessary for a range of downstream biological functions. Additionally this approach fails to consider expression level detail, thus all expressed factors are equally weighted in these assessments.

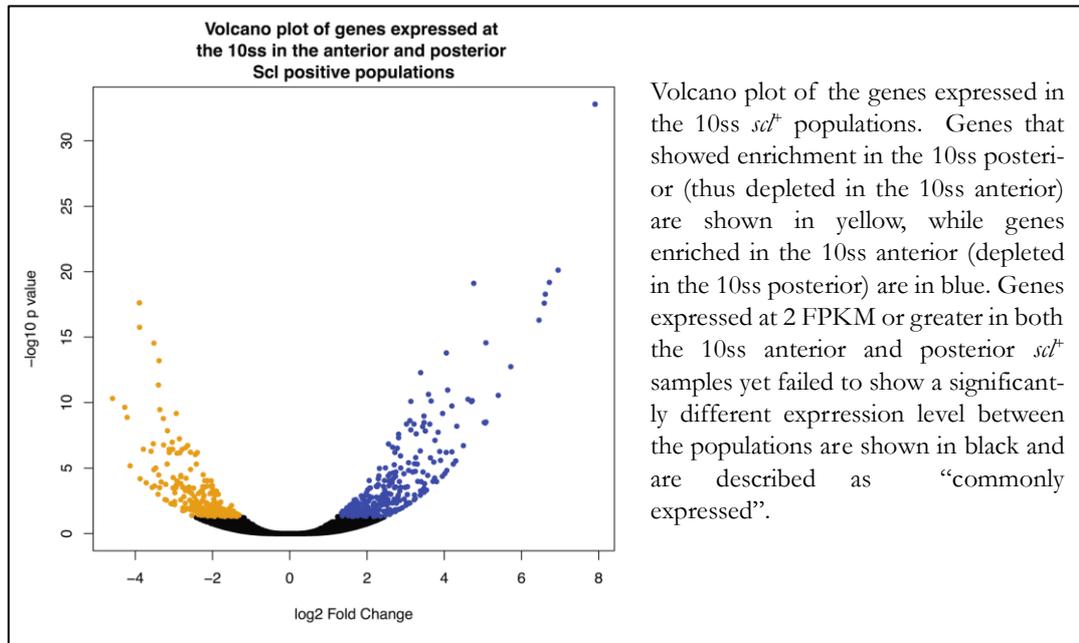
This analysis tool enables large gene sets to be initially probed on a multi-gene level and is a useful approach to identify or validate major biological differences. However due to major caveats as detailed above, this approach requires further gene or population specific studies to validate any biological conclusions drawn. In this project GO enrichment analysis has been used to suggest key biological features of each context.

#### 4.2. Spatial variation in expression at the 10ss.

Mining of these transcriptional datasets for relevant haematopoietic and vascular features is hampered by expression of general factors, which are unrelated to the identity of the cell as part of an *sc<sup>t</sup>* population. Through comparison of two *sc<sup>t</sup>* transcriptomes it is possible to identify genes that are only differentially expressed. Ubiquitous and housekeeping genes are likely to be expressed at similar levels across different populations, thus we can remove many of these using this approach. However factors that are crucial to early haematopoietic development and expressed at similar levels both *sc<sup>t</sup>* contexts, will also be removed by this analysis. This approach is useful for identifying different programs active in different *sc<sup>t</sup>* populations and the potential factors that contribute to these different activities, but it does not resolve whether these are associated with the presence of *sc<sup>t</sup>* or an underlying biological difference. In this initial analysis I have compared the

transcriptomes of the anterior and posterior  $scf^+$  10ss contexts. Figure 4-3 shows the spread and significance of differentially expressed genes between these two populations.

**Figure 4-3** Volcano plot showing the distribution of differentially expressed genes in 10ss  $scf^+$  cells.



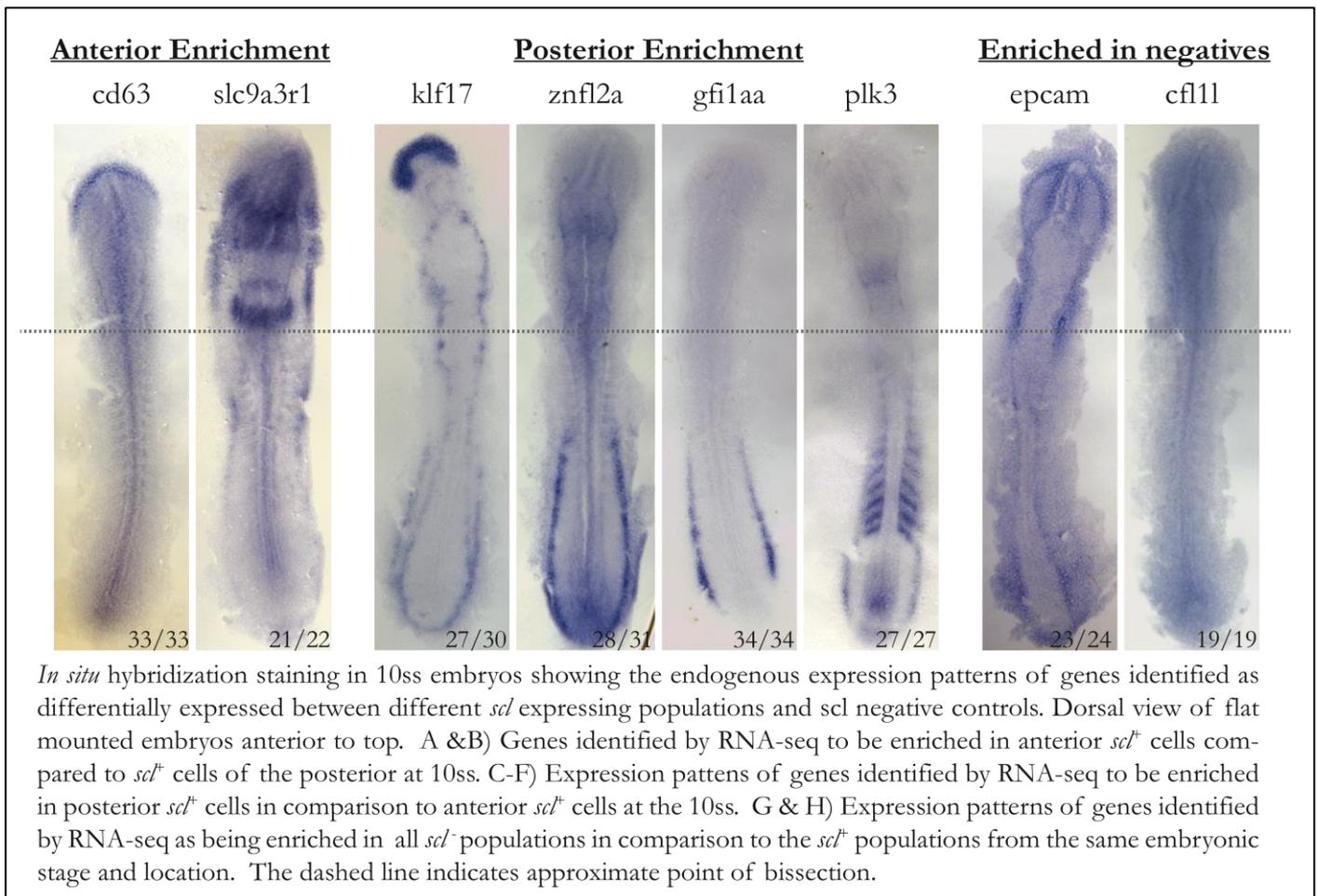
Later in this chapter I use comparative analysis between  $scf^+$  and  $scf^-$  populations to address the specific transcriptional differences between haematopoietic progenitors and the neighbouring cells from within the same biological context (See chapter section 4.5).

#### 4.2.1. Validation of RNA-seq datasets

To validate that RNA-seq datasets accurately represent *in vivo* expression patterns for these early haematopoietic populations I have used whole mount *in situ* hybridization to detect transcript localisation within early zebrafish embryos. DESeq was used to compare expression levels with replicates, between the 10ss anterior and the 10ss posterior  $scf^+$  populations. Target genes for this validation were picked from the most differentially (and significantly) expressed genes showing expression levels of at

least 100 FPKM in one of the two contexts. Many of the most differentially expressed genes are factors that have previously been highly studied in relation to their role in haematopoiesis and vascular development, and their expression pattern in zebrafish has been published for these embryonic stages. Figure 4-4 shows the expression patterns of 8 genes that satisfied these selection criteria in 10ss embryos.

**Figure 4-4 Flat mounted 10ss embryos following *in situ* hybridization for a range of differentially expressed genes identified from RNA-seq datasets produced in this project.**



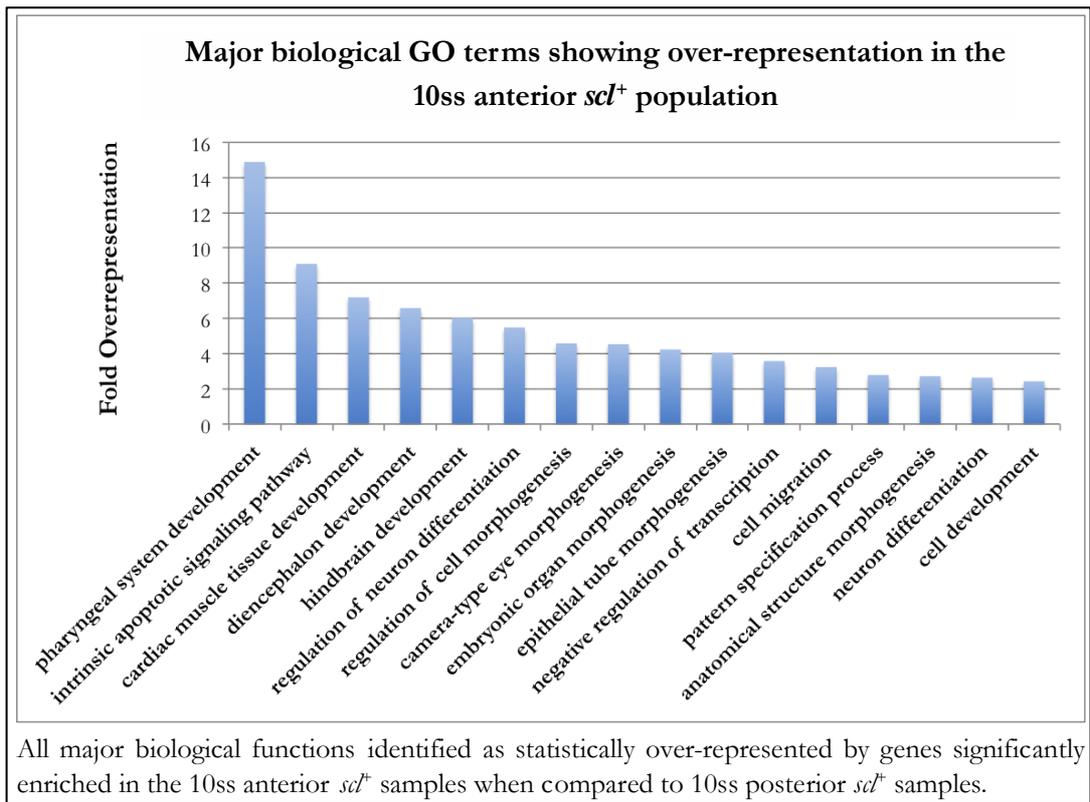
For every gene investigated by *in situ* hybridization I found an expression pattern at the 10ss that matched the RNA-seq datasets, for both *scf*<sup>+</sup> and *scf*<sup>-</sup> contexts. The embryonic location of the *scf*<sup>+</sup> populations profiled by RNA-seq is shown in figure 3-2. It is important to note that my RNA-seq datasets only detail expression levels within the *scf*<sup>+</sup> population for each context, while this embryological technique can

describe expression patterns through out the embryo. Multi-colour *in situ* hybridisation techniques would resolve the degree of overlap between each of these probed factors and the *scf*<sup>+</sup> populations isolated for RNA-seq. Unfortunately standard *in situ* techniques cannot provide the cellular resolution that is required to confirm co-expression with in a single cell. In chapter 6 I detail a range of approaches used to further investigate overlapping expression patterns and potential sub-*scf*<sup>+</sup> context diversity.

#### 4.2.2. Anterior enriched genes

DESeq2 analysis was carried out to compare the transcriptomic datasets for the early (10ss) *scf*<sup>+</sup> anterior and posterior populations. 754 genes were identified as significantly enriched in the anterior. Significant fold enrichment varied from 2.5-240x over expression in the posterior at the 10ss. Top over-represented biological processes were determined using the PANTHER database, revealing that 93 terms were significantly over-represented. Unlike in the temporal comparison of the posterior the majority of these top enriched terms relate to tissue-specific biological functions, including kidney, neural and eye development (figure 4-5). No haematopoietic or vascular processes are found among the significantly over-represented biological processes despite the expression of key haemangiogenic factors including *scf* in both of the populations. These results highlight that comparison between *scf*<sup>+</sup> populations can identify differences regardless of whether these are specific to the *scf* program or a result of underlying differences, which can swamp the analysis

Figure 4-5 GO enrichment analysis of genes enriched in the 10ss anterior *scf*<sup>+</sup> population compared to the 10ss posterior *scf*<sup>+</sup> population.



The top featuring protein class was transcription factors at 11.5%, while 8.5% of proteins were signalling molecules. The genes annotated as transcription factors relate to 141 over-represented biological GO processes including nervous system development and cardiac muscle cell differentiation. The genes annotated as signalling molecules correlate with 57 significantly over-represented biological processes, which include Wnt signalling (canonical and non-canonical) and axon guidance. Although these gene lists may contain interesting novel factors involved in anterior haematopoietic and vascular development, these will be difficult to identify from the large number of genes included that are involved in other anterior cell processes.

The lack of significant over-representation of haematopoietic or vascular GO terms in the 10ss anterior enriched gene list could be a result of a variety of biological situations.

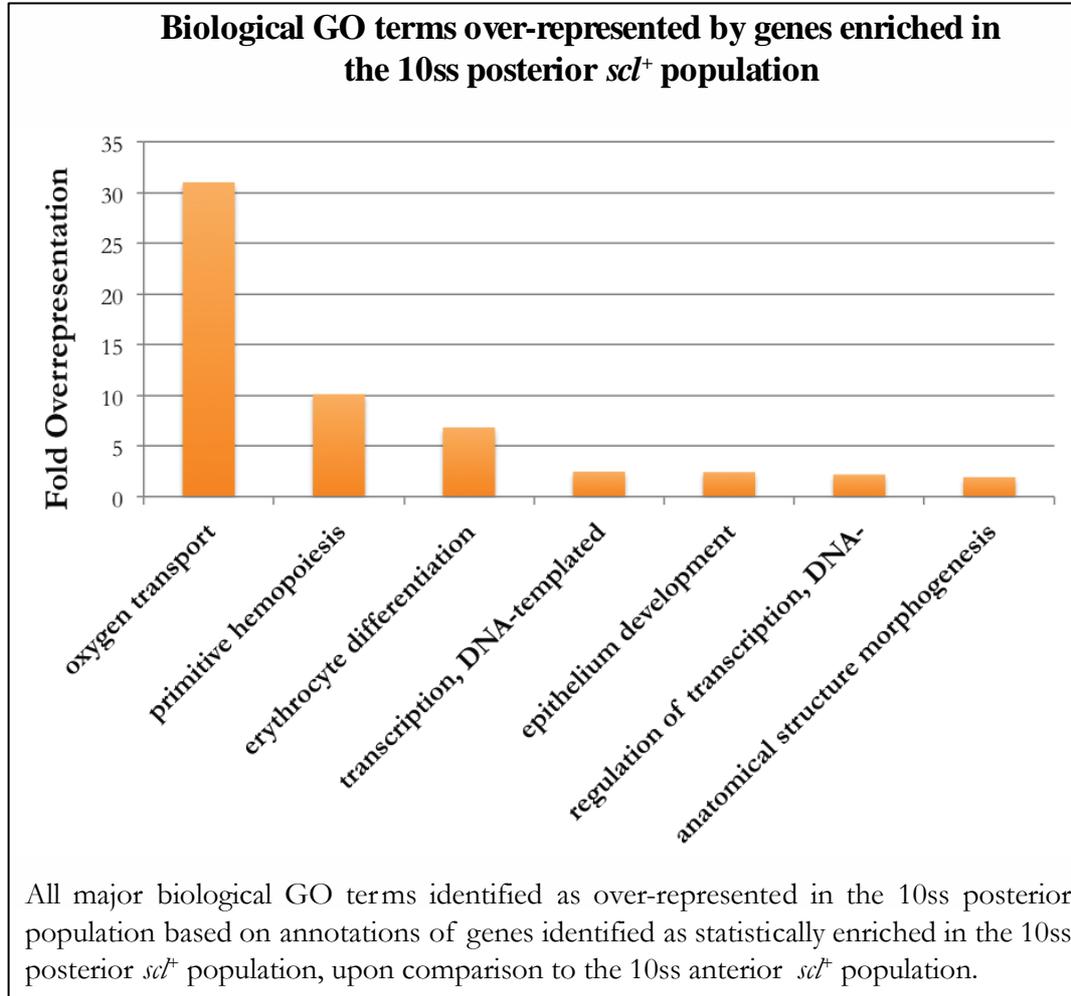
- The 10ss anterior *sc<sup>+</sup>* population may show enriched haematopoietic and vascular expression compared to its neighbouring *sc<sup>-</sup>* cells. The majority of these factors may be expressed at higher levels in the 10ss posterior *sc<sup>+</sup>* population, thus upon comparison to the posterior rather than neighbouring *sc<sup>-</sup>* cells, the 10ss anterior appears to be depleted in the very factors that are key to its cellular activity.
- The 10ss anterior *sc<sup>+</sup>* population may be highly diverse and still in the process of specification, thus shows expression of factors involved in the cellular processes of neighbouring *sc<sup>-</sup>* cell types. Later in development these genes are down regulated in *sc<sup>+</sup>* cells and become restricted to their specific subpopulations. This range of genes expressed in the 10ss anterior relating to a large number of biological processes, reduces the significance of genes being specifically expressed in the 10ss anterior population.

#### 4.2.3. Posterior enriched genes

744 genes are significantly enriched in the 10ss posterior (thus depleted in the 10ss anterior). From this posterior enriched gene list 71 biological GO terms are significantly enriched. The top two major terms that are over-represented are for oxygen transport and primitive haematopoiesis, suggesting that the posterior is already adopting the erythroid fate (figure 4-6). This is very different to the anterior enriched gene list, which showed no significant haematopoietic process over-representation. The 33 genes that contribute to these biological terms are highly expressed with a mean FPKM over 140. Table 4-1 details the expression level and enrichment of these erythroid factors enriched in the posterior in this 10ss spatial transcriptome comparison. Other over-represented biological functions relate to transcription and general regulation of gene expression. The top enriched molecular

GO functions are also strongly erythroid including oxygen transporters, oxygen binding and iron ion binding.

Figure 4-6 Biological and protein class analysis of genes enriched in the 10ss posterior *scl*<sup>+</sup> expressing population.



Protein class analysis shows that 11.6% of the genes enriched in the posterior *scl*<sup>+</sup> population are annotated as transcription factors and that 5.4% are signalling molecules. Top over-represented biological GO terms for the 10ss posterior enriched transcription factor class include “primitive hemopoiesis” (associated with enriched 10ss posterior *etv5a*, *gfi1aa*, *gata1a* and *cebpa* expression) and the vascular development.

**Table 4-1 Erythroid factors enriched in the 10ss posterior *scf* expressing population.**

Gene Name	Description	10ss Ant. Mean FPKM	10ss Post. Mean FPKM	log2 FoldChange	p-value
hbac1	hemoglobin, alpha embryonic 1	0.93	19.47	-3.4697	6.13E-08
gfi1aa	growth factor independent 1A transcription repressor a	148.74	1861.89	-3.4001	4.65E-15
gata1a	GATA binding protein 1a	50.29	634.37	-3.3634	6.15E-13
hbac3	hemoglobin alpha embryonic-3	33.60	407.10	-3.2744	3.51E-12
hbbe1.1	hemoglobin beta embryonic-1.1	16.51	216.93	-3.2663	6.39E-10
hbz	hemoglobin zeta	0.45	8.70	-3.2107	1.69E-06
csnrp1a	cysteine-serine-rich nuclear protein 1a	2.35	34.02	-3.1757	2.78E-08
hbbe3	hemoglobin beta embryonic-3	30.11	348.91	-3.1224	1.54E-09
hbac1	hemoglobin, alpha embryonic 1	13.44	132.99	-2.7685	2.95E-06
mb	myoglobin	0.68	7.97	-2.7400	2.20E-05
hbbe2	hemoglobin beta embryonic-2	4.63	43.52	-2.4164	0.00464013
wnt16	wingless-type MMTV integration site family, member 16	0.36	3.72	-2.3010	0.001885645
mta3	metastasis associated 1 family, member 3	8.36	45.97	-2.2492	2.69E-06
hbac1	hemoglobin, alpha embryonic 1	0.86	7.54	-2.2056	0.002087583
flvcr1	feline leukemia virus subgroup C cellular receptor 1	4.43	21.13	-2.0646	1.22E-05
hbbe1.2	hemoglobin beta embryonic-1.2	0.48	5.89	-2.0542	0.008482521
epb41b	erythrocyte membrane protein band 4.1b	1.30	7.42	-1.8827	0.005080177
hbbe1.1	hemoglobin beta embryonic-1.1	8.76	44.15	-1.8575	0.004529907
slc4a1a	solute carrier family 4 (anion exchanger), member 1a	0.49	2.32	-1.6862	0.01238688
tbx16	T-box 16	3.68	14.48	-1.6300	0.004668324
fech	ferrochelatase	7.27	25.91	-1.5010	0.013413892
HBZ	zgc:163057	2.96	12.24	-1.4951	0.029360255
etv5a	ets variant 5a	9.49	28.57	-1.4538	0.001603395
pms2	PMS1 homolog 2, mismatch repair system component	4.29	13.18	-1.4060	0.004378818
cebpa	CCAAT/enhancer binding protein (C/EBP), alpha	107.65	300.95	-1.3700	0.001766542
alas2	aminolevulinic acid, delta-, synthase 2	10.79	34.11	-1.3647	0.013860034
myb	v-myb avian myeloblastosis viral oncogene homolog	52.87	141.33	-1.3021	0.002925659
slc25a37	solute carrier family 25 (mitochondrial iron transporter)	5.69	15.25	-1.1924	0.042964406
smox	spermine oxidase	10.43	25.62	-1.1776	0.009596979
cdx4	caudal type homeobox 4	40.14	96.74	-1.0939	0.025151866
gata2a	GATA binding protein 2a	11.94	25.33	-0.9993	0.019734823
piezo1	piezo-type mechanosensitive ion channel component 1	7.18	15.39	-0.9954	0.029586497
cdca7a	cell division cycle associated 7a	72.29	152.12	-0.9863	0.015035026

Table of erythroid factors identified through biological gene ontology analysis of genes enriched in the 10ss posterior *scf*<sup>+</sup> population upon comparison to expression in the 10ss anterior *scf*<sup>+</sup> population. Factors are in order of fold depletion in the 10ss anterior compared to the expression level in the 10ss posterior, thus are negative log2 values. This gene list is biologically equivalent to genes enriched in the 10ss posterior.

For genes annotated as signalling molecules over-represented biological processes relate to general developmental processes such as limb development and dorsal/ventral patterning.

This spatial comparison of *scf* populations indicates the posterior 10ss *scf* transcriptome is already strongly enriched for erythroid factors while the 10ss anterior population shows no significant over-representation of any haematopoietic related GO term.

#### 4.2.4. Common expressed genes in the 10ss spatial *scf*<sup>+</sup> comparison

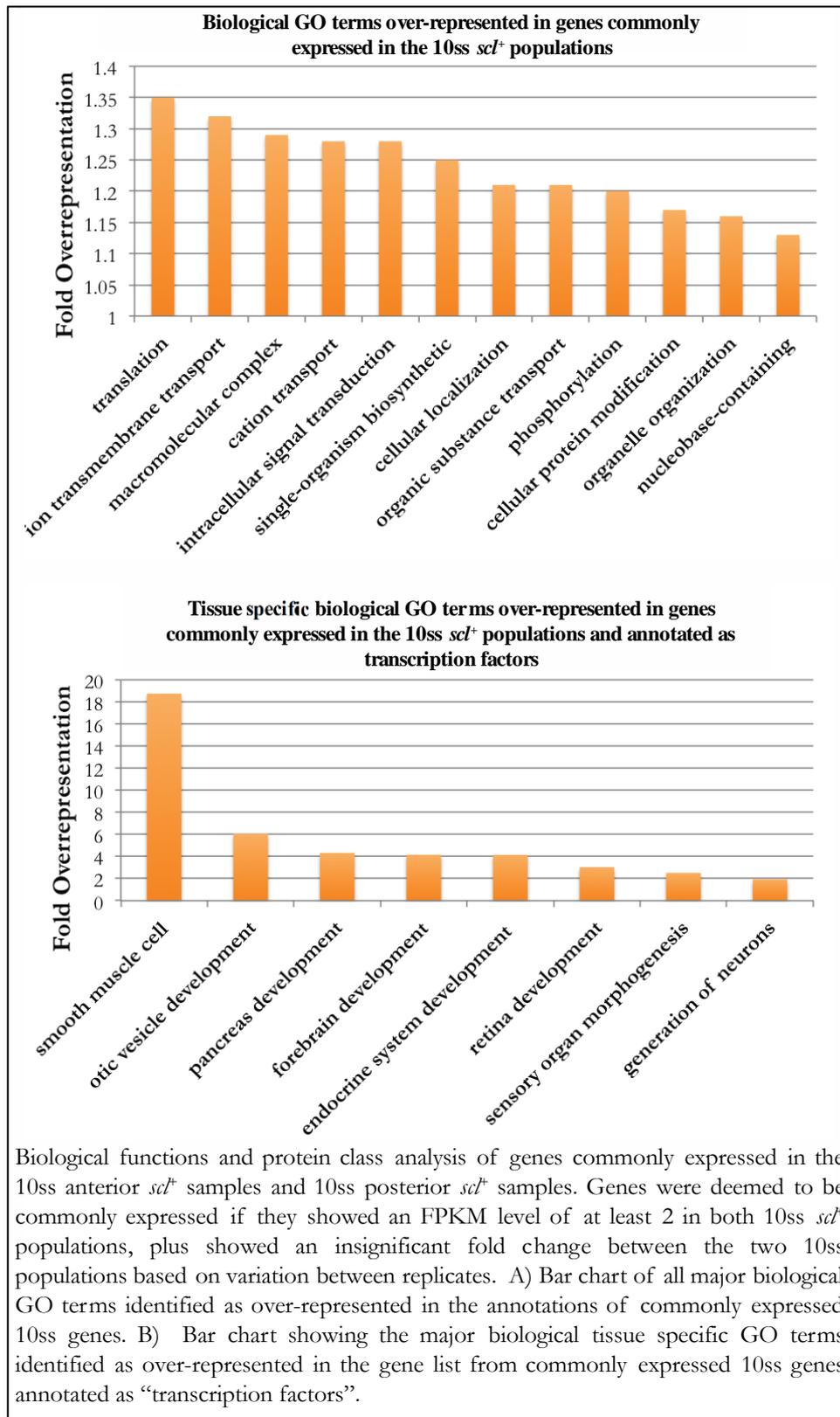
As with the temporal comparison of the posterior genes, genes with FPKM values greater than 2 in both the 10ss *scf*<sup>+</sup> populations (anterior and posterior) without significant ( $p > 0.05$ ) difference in expression level between the two contexts, were considered commonly expressed. 8991 genes for this spatial comparison were commonly expressed and annotated with 310 over-represented biological GO terms showing significant difference to that expected if genes were picked at random: 259 were enriched and 51 were depleted.

All but 6 of these biological terms were for ubiquitous housekeeping processes or general development (figure 4-7). The tissue-specific categories relate to retina development and myeloid cell differentiation. This over-representation could suggest that at the 10ss the myeloid fate is favoured by both the anterior and posterior *scf*<sup>+</sup> populations, however the common expression of factors annotated with a role in retina development casts doubt on the validity of such a conclusion.

8.5% of commonly expressed posterior factors were classed as transcription factors by the PANTHER database. The commonly expressed transcription factors relate to a variety of ubiquitous biological processes plus brain and sensory organ development.

3.2% of enriched genes were annotated as signalling molecules, these related to key signal transduction pathways, rhombomere formation and vascular development. Non-canonical Wnt, SMAD and BMP signalling pathways were enriched, as were negative regulators of Ras signalling and the Jak/STAT cascade.

Figure 4-7 Biological gene ontology and protein class analysis of genes expressed at the 10ss in both the anterior and posterior *scf* expressing populations.

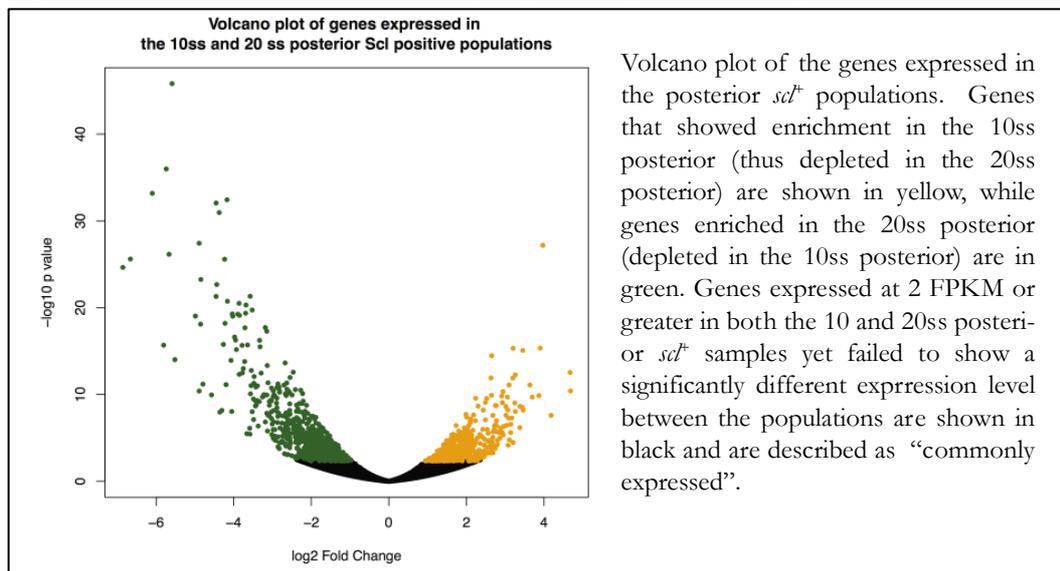


### 4.3. Temporal transcriptome variation in the posterior- 10ss vs. 20ss

#### 4.3.1. 10ss enrichment in *scf*<sup>+</sup> posterior comparison

Following DESeq2 analysis comparing the transcriptomic datasets for the *scf*<sup>+</sup> posterior 10 and 20ss samples, 471 genes were identified as significantly enriched in the 10ss posterior. The spread and significance of differential expression is shown in figure 4-8. Significant fold enrichment varied from 1.9- 25.7 fold over expression in the 20ss posterior.

**Figure 4-8 Distribution of differentially expressed genes in *scf*<sup>+</sup> posterior cells.**

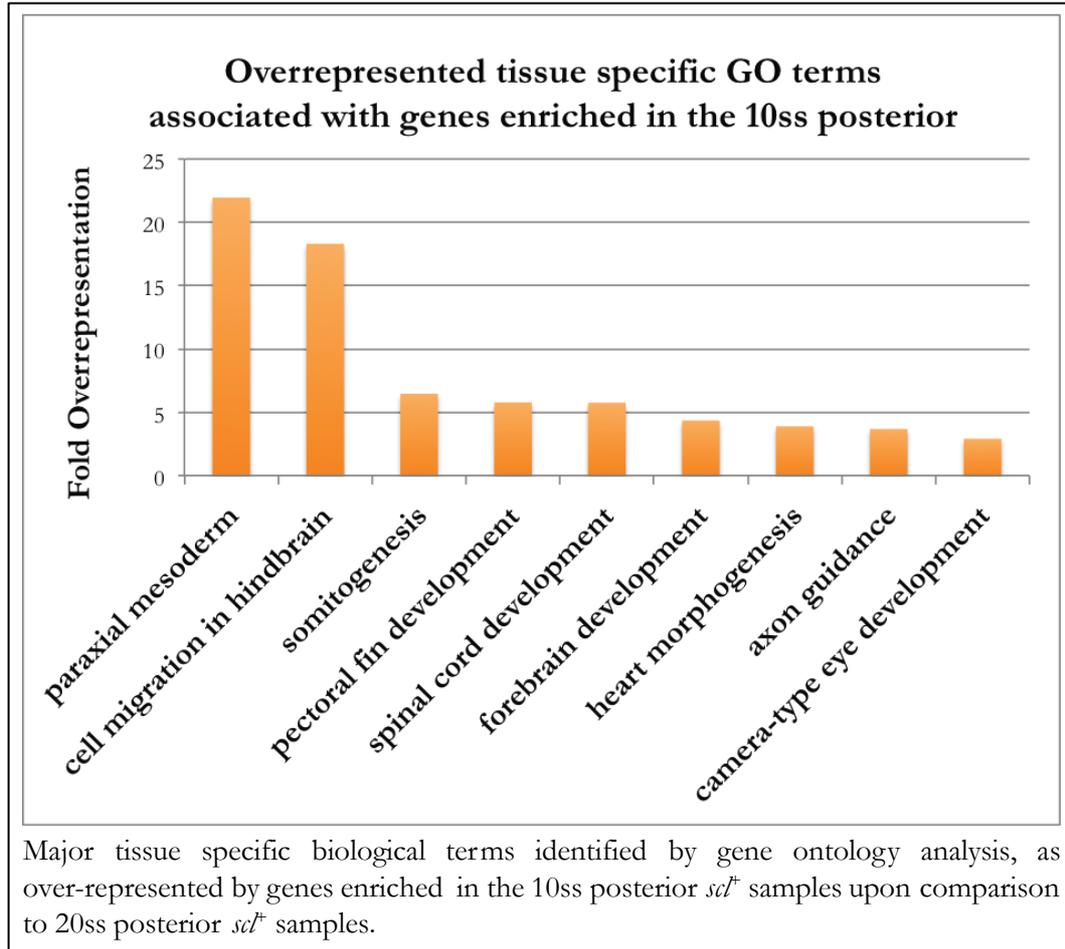


Top enriched biological processes were determined using the PANTHER database, revealing that 67 terms were over-represented in the 10ss posterior compared to the 20ss posterior, all of which relate to ubiquitous processes apart from 9 (see figure 4-9).

Of these 9 none are directly haematopoietic or vascular, suggesting that all cell type specific biological processes active in the 10ss posterior transcriptome are maintained in the 20ss posterior cells. The expression of the genes that contribute to these terms are reduced in the 20ss posterior (data not shown) suggesting that the

posterior *sc<sup>+</sup>* population has become more specialised in function, as it develops from the 10 to 20ss.

**Figure 4-9 Biological gene ontology analysis genes enriched in the 20ss posterior *sc<sup>+</sup>* population**



Protein class analysis annotates 11.6% of the 10ss enriched genes in this posterior comparison, as transcription factors and 6.5% are annotated as signalling molecules. The genes annotated as transcription factors relate to 107 biological GO processes including development of the kidneys, nervous system and myeloid cell differentiation. Expression of the majority of genes annotated with these processes has dropped in the 20ss *sc<sup>+</sup>* posterior possibly because the *sc<sup>+</sup>* population has become further specified.

Genes for transcription factors enriched in the early posterior annotated with a role in myeloid cell differentiation are also over-represented suggesting that the

development of the myeloid lineage has either dramatically reduced or migrated to the anterior by the 20ss. However as seen in section 4.5.5 genes associated with myeloid cell differentiation are specifically enriched in the 20ss posterior *sc<sup>l+</sup>* population (over *sc<sup>l-</sup>* cells), suggesting that this process remains active but at a lower level than in the 10ss *sc<sup>l+</sup>* posterior population. The genes annotated as signalling molecules correlate with 82 enriched biological processes, which include regulation of signal transduction, eye development and myeloid cell development. This correlates with the population retaining haematopoietic potential apart from the myeloid lineage, as the posterior *sc<sup>l+</sup>* cells develop, however it is important to remember that this analysis is entirely based on the accuracy of each gene's annotation.

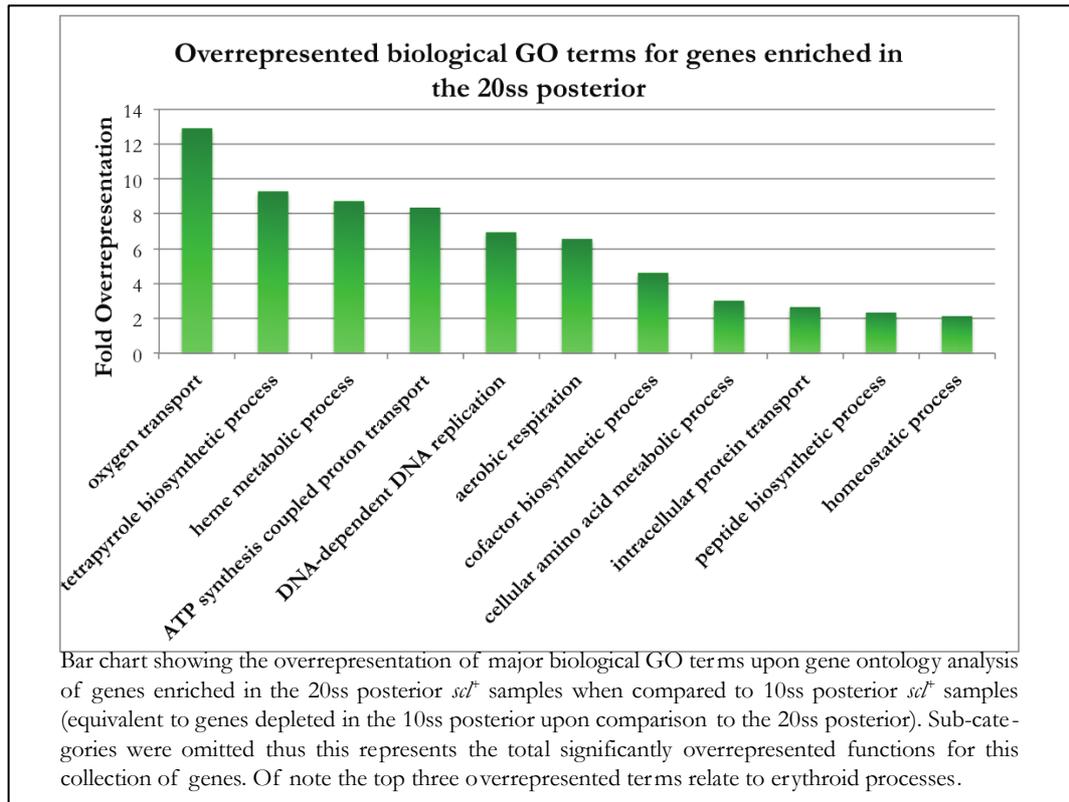
In conclusion 10ss enriched factors in the posterior comparison of *sc<sup>l+</sup>* cells' transcriptomes are annotated with non-haemangiogenic functions. This suggests all haematopoietic functions expressed at the 10ss are retained to the 20ss in the posterior of the zebrafish embryo. However upon focussing solely on transcription factor genes enriched in the 10ss posterior significant over-representation was observed for several biological functions including myeloid differentiation.

#### 4.3.2. 20ss enrichment in *sc<sup>l+</sup>* posterior comparison

In this temporal comparison of the posterior *sc<sup>l+</sup>* cells 1564 genes are significantly enriched in the 20ss posterior (thus depleted in the 10ss posterior). A greater number of genes are enriched in the 20ss sample than in the 10ss transcriptome, which may be due to the increasing complexity and heterogeneity of the population.

From the 20ss enriched gene list only 92 biological GO terms are significantly over-represented, these are shown in figure 4-10.

Figure 4-10 Biological gene ontology analysis genes enriched in the 20ss posterior *scf*<sup>+</sup> population.



The most significantly enriched biological process associated with these 20ss enriched factors is oxygen transport. Unlike in the anterior temporal comparisons, this magnitude of over-representation in biological function is much greater and relates to 25 genes expressed with a mean FPKM of over 1200. Similarly the top enriched molecular GO functions are oxygen transport, oxygen binding and iron ion binding. This analysis suggests that progression of the posterior *scf*<sup>+</sup> population from the 10ss to 20ss involves a dramatic shift towards a strongly erythroid profile, without any notable divergence or range of function. Table 4-2 details the expression level and enrichment of these oxygen transport factors enriched in the 20ss *scf*<sup>+</sup> posterior population compared to the 10ss posterior population.

**Table 4-2 Genes associated with the GO term oxygen transport enriched in the 20ss posterior *scf<sup>+</sup>* population**

Gene Name	Description	10ss Post. Mean FPKM	20ss Post. Mean FPKM	log2 FoldChange	pvalue
hbbe1.1	hemoglobin beta embryonic-1.1	44.15	2417.19	-5.590697375	1.54E-46
hbbe1.2	hemoglobin beta embryonic-1.2	5.89	211.58	-4.890837669	3.77E-28
hbbe3	hemoglobin beta embryonic-3	348.91	8086.59	-4.454357537	8.83E-33
hbbe2	hemoglobin beta embryonic-2	43.52	1078.15	-4.436969149	2.13E-23
hbae1	hemoglobin, alpha embryonic 1	19.47	431.08	-4.374632586	1.10E-31
hbae1	hemoglobin, alpha embryonic 1	132.99	2678.09	-4.234442923	2.63E-26
hbae3	hemoglobin alpha embryonic-3	407.10	7583.52	-4.172158985	3.71E-33
HBZ	zgc:163057	12.24	215.73	-3.952978035	5.36E-17
hmbsb	hydroxymethylbilane synthase, b	16.42	262.62	-3.893263409	5.56E-20
hbz	hemoglobin zeta	8.70	123.01	-3.709571876	2.07E-18
hbbe1.1	hemoglobin beta embryonic-1.1	216.93	2911.75	-3.682892629	4.91E-21
urod	uroporphyrinogen decarboxylase	161.98	1958.75	-3.57508809	4.86E-22
alas2	aminolevulinate, delta-, synthase 2	34.11	352.61	-3.221007282	1.04E-10
hbae1	hemoglobin, alpha embryonic 1	7.54	63.21	-2.958728432	1.20E-10
hmox1a	heme oxygenase 1a	4.29	25.45	-2.441015132	1.12E-06
uros	uroporphyrinogen III synthase	15.01	76.89	-2.379187583	1.68E-09
cpox	coproporphyrinogen oxidase	216.83	1010.97	-2.27233994	3.21E-11
fech	ferrochelatase	25.91	122.42	-2.243091816	9.64E-07
hmbsa	hydroxymethylbilane synthase a	104.51	440.51	-2.12380162	1.78E-09
iba57	IBA57 homolog, iron-sulfur cluster assembly	17.20	52.31	-1.675982817	9.03E-06
snx3	sorting nexin 3	13.66	37.67	-1.515227165	0.000112343
atp11b	ATPase inhibitory factor 1b	46.07	123.56	-1.513920269	9.62E-05
bdh2	3-hydroxybutyrate dehydrogenase, type 2	3.41	7.03	-1.088473712	0.048504856
blvra	biliverdin reductase A	72.48	127.73	-0.951309948	0.023301985

Table of factors associated with erythroid biological GO terms, identified as enriched genes in the 20ss posterior *scf<sup>+</sup>* population following transcriptome comparison with samples for the *scf<sup>+</sup>* 10ss posterior. Biological GO analysis was carried out on the total enriched gene list regardless of protein class or molecular functionality. Factors are in order of fold depletion in the 10ss posterior compared to the expression level in the 20ss posterior, thus are negative log2 values. This gene list is biologically equivalent to genes enriched in the 20ss posterior. Gene name and description provided by the Ensemble biomart database.

Protein class analysis shows that 4.6% of the genes enriched in the 20ss *scf<sup>+</sup>* posterior population (in this temporal comparison) are annotated as transcription factors and that 2.3% are signalling molecules. Top protein classes instead include hydrolases, transferases and transporters, supporting previous reports that by 20ss the erythroid population has been differentiated and gains functionality. Transcriptional regulation, rather than a tissue-specific function, was the most commonly over-represented biological GO terms relating to the transcription factor class. Notably, no over-represented terms relating to tissue-specific processes or regulation of a developmental process were identified from the 20ss posterior enriched transcription factor list.

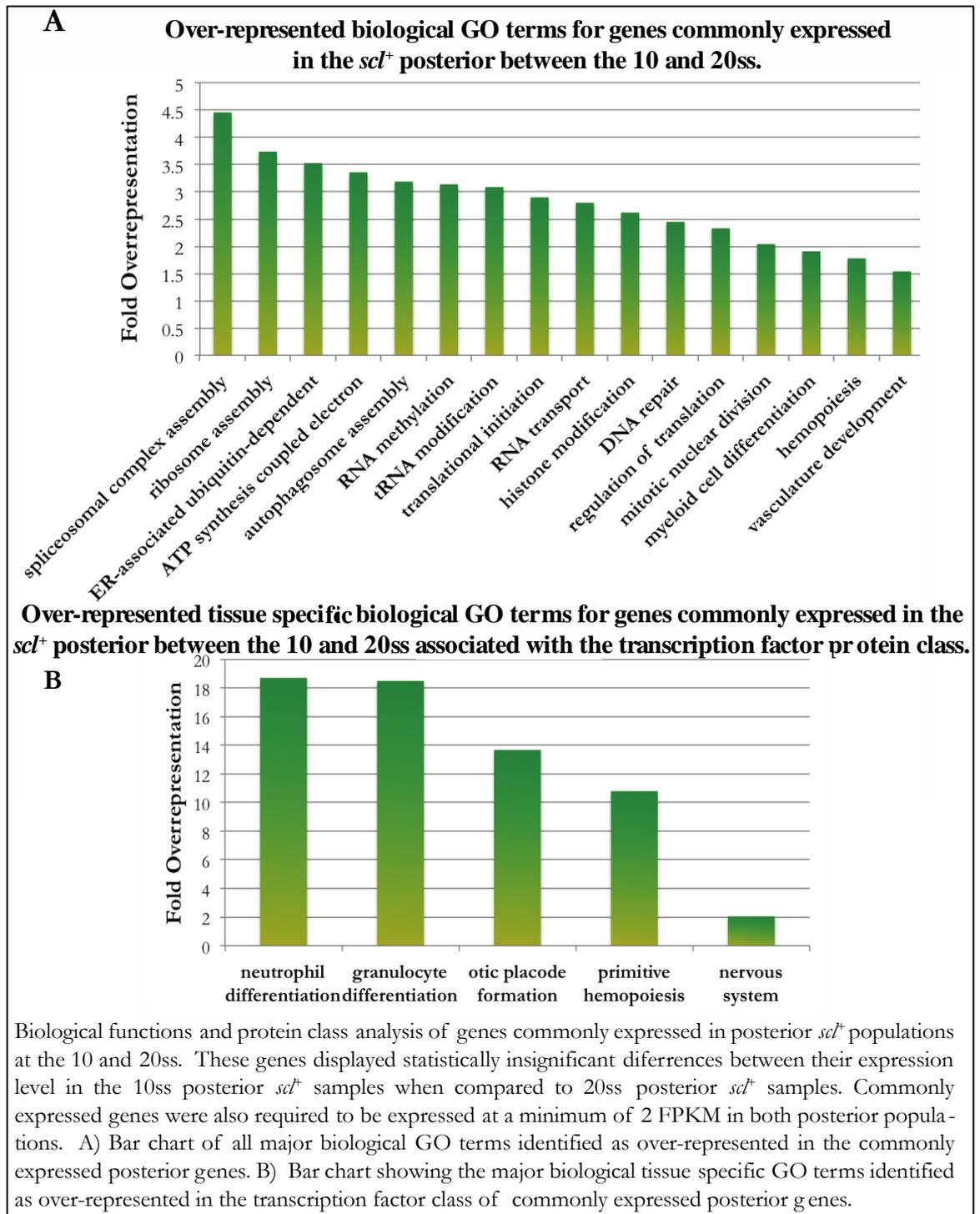
### 4.3.3. Common factors from the *sc<sup>l+</sup>* posterior comparison

Genes with FPKM values greater than 2 in both the 10ss and 20ss *sc<sup>l+</sup>* posterior cells without significant ( $p > 0.05$ ) difference in expression level between the two contexts, were considered commonly expressed. 7656 genes for this posterior comparison were commonly expressed and annotated with 253 biological GO terms. All but 3 of these terms were for ubiquitous housekeeping processes or general development. The 3 tissue-specific categories were all haematopoietic or vascular processes, as detailed in figure 4-11. The commonly expressed genes that contribute to the over-representation of these haematopoietic or vascular processes indicates that despite the erythroid fate becoming strongly favoured as the population develops, the expression of certain general haematopoietic factors and vascular genes was maintained.

9.9% of commonly expressed posterior factors were classed as transcription factors by the PANTHER database. Interestingly the commonly expressed transcription factors relate to a variety of biological processes including otic vesicle development, myeloid and leukocyte differentiation in addition to ubiquitous processes such as circadian rhythm. 3.5% of enriched genes were annotated as signalling molecules, these related to both key signal transduction pathways and haemangiogenic signalling. Non-canonical Wnt signalling and pathways involving SMAD proteins were enriched as were negative regulators of Ras signalling. This correlates with previous studies that show that Wnt signalling is important to the erythroid fate, but without validation within these developmental contexts, this is only a correlation<sup>263-</sup>

<sup>265</sup>.

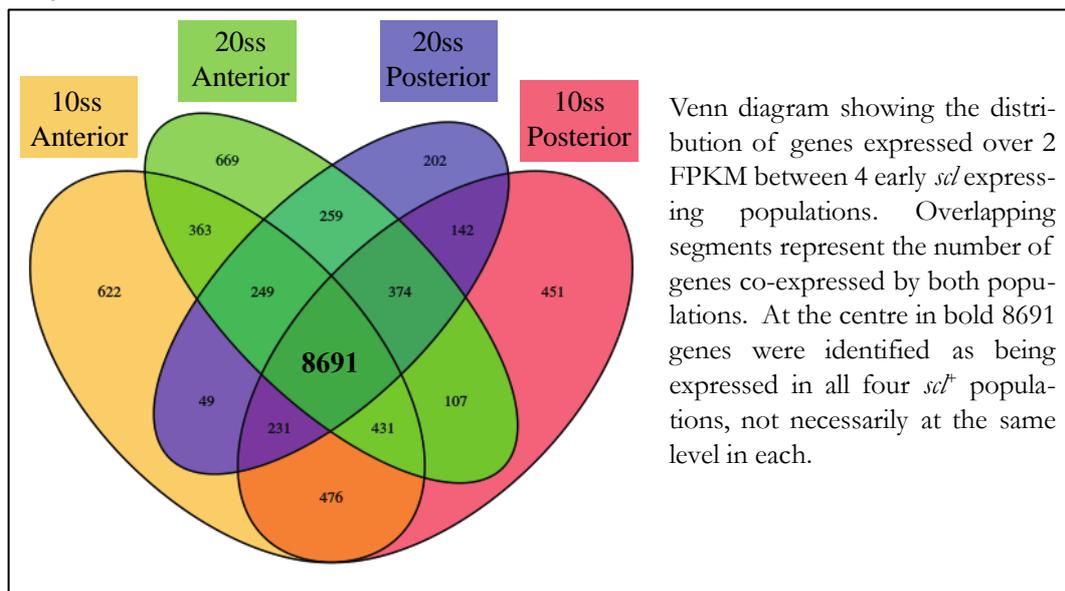
Figure 4-11 Protein class and gene ontology analysis of genes commonly expressed in the 10 and 20ss posterior *scl*<sup>+</sup> populations.



#### 4.4. Common *scf*<sup>+</sup> cell factors

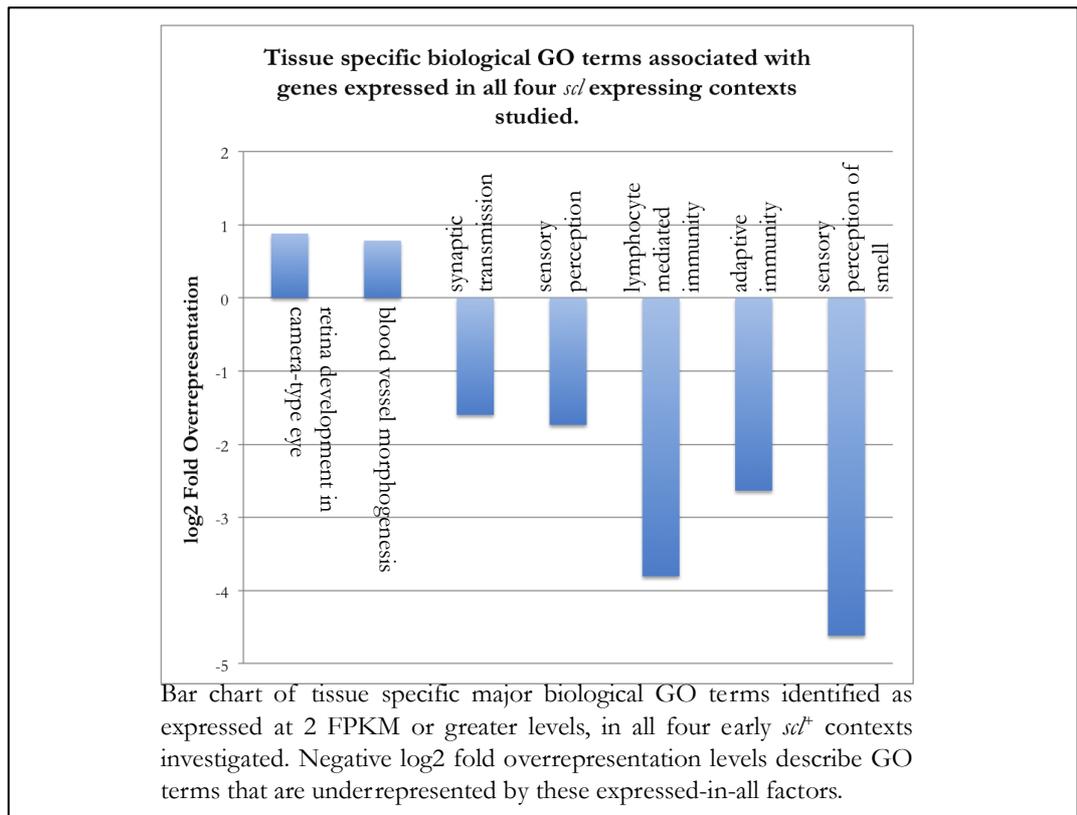
8691 genes were reproducibly expressed in all contexts with a minimal FPKM of 2; their distribution between the four *scf*<sup>+</sup> contexts is shown in figure 4-12. This gene list included factors that showed varied expression levels between the different contexts.

Figure 4-12 Venn diagram showing the distribution of genes expressed in all four early *scf*<sup>+</sup> contexts.



Biological process enrichment analysis showed that 332 GO terms common to these four *scf* populations were over-represented and 30 terms were underrepresented in this common gene list. 7 of the over-represented processes were for tissue-specific functions, falling into two major biological GO terms, retina development and blood vessel morphogenesis as detailed in Figure 4-13. 47 ubiquitous major biological GO terms were also over-represented- RNA processing and ATP synthesis processes accounted for the top over-represented ubiquitous biological GO terms.

Figure 4-13 Biological gene ontology analysis of genes expressed by all four *scf*<sup>+</sup> populations studied.



Within the depleted GO terms 5 were for tissue-specific processes such as sensory perception, synaptic transmission and immune responses, which indicates that these biological function are commonly absent and potentially specifically repressed within *scf*<sup>+</sup> expressing cells. Depletion within this common list could also be interpreted that these processes are enriched within certain *scf*<sup>+</sup>-subpopulations, while being completely absent in others, thus would fail to be included in this “expressed-in-all” analysis.

Upon protein class analysis of these commonly expressed proteins, transcription factors account for 9.1% of the genes and 3.5% are annotated as signalling molecules. Over-represented tissue-specific biological processes for the common expressed genes classed as transcription factors included granulocyte differentiation, primitive haematopoiesis and blood vessel morphogenesis (genes listed in table 4-3).

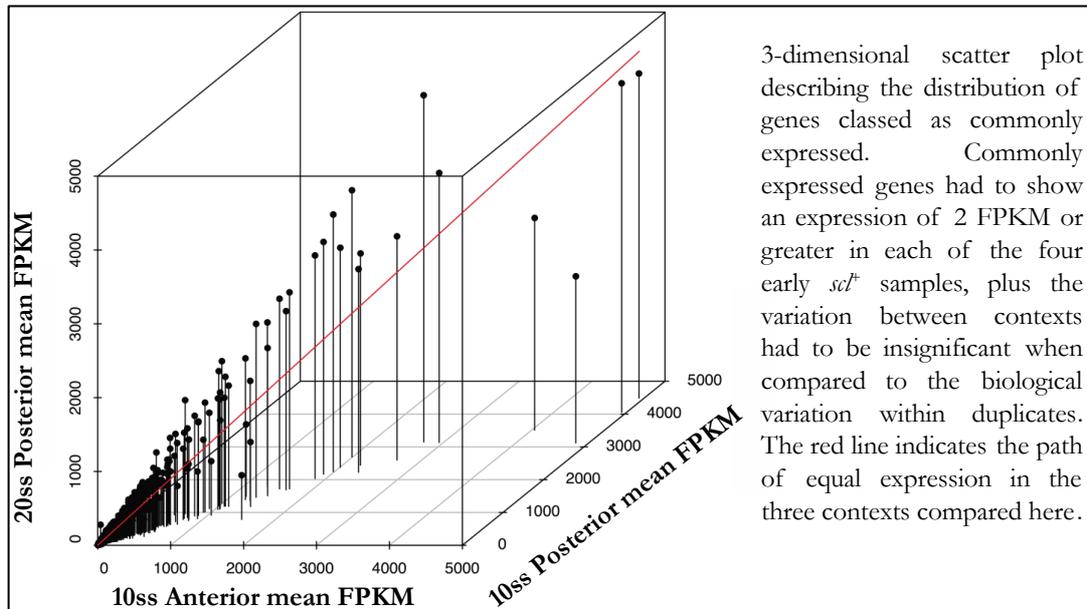
**Table 4-3 Genes associated with a molecular GO term of transcription factor and a biological GO term of haematopoiesis expressed by all four *scf*<sup>+</sup> populations studied.**

Gene Name	Description	10ss Ant.	10ss Post.	20ss Ant.	20ss Post.
		Mean FPKM	Mean FPKM	Mean FPKM	Mean FPKM
ajuba	ajuba LIM protein	15.90	18.17	29.07	17.31
bcl6a	B-cell CLL/lymphoma 6a (zinc finger protein 51)	8.71	10.34	19.95	11.78
bcl6ab	B-cell CLL/lymphoma 6a, genome duplicate b	16.92	10.93	32.57	15.56
brf1a	BRF1, RNA polymerase III transcription initiation factor a	3.31	3.71	3.90	2.02
brf1b	BRF1, RNA polymerase III transcription initiation factor b	11.12	8.15	9.35	9.20
cdx4	caudal type homeobox 4	40.14	96.74	4.88	26.79
cebpa	CCAAT/enhancer binding protein (C/EBP), alpha	107.65	300.95	208.21	168.40
crip2	cysteine-rich protein 2	32.78	26.92	325.02	110.04
dpf3	D4, zinc and double PHD fingers, family 3	5.85	2.18	6.73	3.79
e2f7	E2F transcription factor 7	5.19	8.13	11.06	10.46
e2f8	E2F transcription factor 8	7.64	12.99	9.31	14.54
ell	elongation factor RNA polymerase II	14.91	14.56	15.98	12.81
erg	v-ets avian erythroblastosis virus E26 oncogene homolog	10.69	10.36	86.16	16.02
etsv2	ets variant 2	1535.57	788.51	1384.75	598.63
etsv5a	ets variant 5a	9.49	28.57	33.20	18.12
fev	FEV (ETS oncogene family)	4.42	5.18	3.73	3.40
fli1a	Fli-1 proto-oncogene, ETS transcription factor a	151.39	304.34	459.33	281.84
fli1b	Fli-1 proto-oncogene, ETS transcription factor b	65.15	43.94	222.47	69.70
fosab	v-fos FBJ murine osteosarcoma viral oncogene homolog Ab	1437.93	1182.45	1003.04	873.22
foxc1a	forkhead box C1a	6.50	2.85	14.80	2.32
foxb1b	forkhead box J1b	3.62	6.76	2.11	13.64
foxo3b	forkhead box O3b	2.29	2.95	6.09	4.00
gata1a	GATA binding protein 1a	50.29	634.37	50.12	470.48
gata2a	GATA binding protein 2a	11.94	25.33	31.02	14.70
gata3	GATA binding protein 3	33.34	28.45	23.78	15.91
gata5	GATA binding protein 5	146.72	9.41	28.95	2.25
gf11aa	growth factor independent 1A transcription repressor a	148.74	1861.89	90.34	793.09
hey2	hes-related family bHLH transcription factor with YRPW motif 2	100.58	78.62	164.85	69.28
ikzf1	IKAROS family zinc finger 1 (Ikaros)	13.70	11.77	71.73	106.80
kdm4ab	lysine (K)-specific demethylase 4A, genome duplicate b	5.89	7.45	8.18	6.88
lmo2	LIM domain only 2 (thombotin-like 1)	3278.00	3763.49	1312.93	1438.30
mafba	v-maf avian musculoaponeurotic fibrosarcoma oncogene Ba	6.84	5.73	13.07	4.74
med14	mediator complex subunit 14	5.25	8.80	8.20	8.05
meox1	mesenchyme homeobox 1	31.39	24.71	33.37	9.39
mgaa	MGA, MAX dimerization protein a	11.05	13.12	4.68	5.11
myb	v-myb avian myeloblastosis viral oncogene homolog	52.87	141.33	70.02	165.47
ncor1	nuclear receptor corepressor 1	41.81	24.20	17.33	11.88
ncor2	nuclear receptor corepressor 2	14.93	11.65	22.29	11.77
nfil3	nuclear factor, interleukin 3 regulated	22.29	11.32	34.36	11.38
nr2f2	nuclear receptor subfamily 2, group F, member 2	4.23	2.96	15.13	6.12
osr1	odd-skipped related transcription factor 1	6.55	26.09	2.59	5.69
pdcd2	programmed cell death 2	19.88	20.41	16.00	34.32
prox1b	prospero homeobox 1b	6.66	16.38	10.37	6.52
rarb	retinoic acid receptor, alpha b	11.00	7.22	6.24	4.19
rbpa	recombination signal binding protein for Ig kappa J region a	8.61	12.75	12.03	7.52
rbpb	recombination signal binding protein for Ig kappa J region b	6.89	6.89	11.20	6.33
sbds	Shwachman-Bodian-Diamond syndrome	22.72	24.11	16.08	23.43
smad5	SMAD family member 5	83.31	103.78	127.93	75.49
smad6a	SMAD family member 6a	10.29	3.85	8.20	2.40
smad9	SMAD family member 9	18.24	7.94	7.60	4.50
smyd3	SET and MYND domain containing 3	9.57	6.10	9.22	16.43
sox18	SRY (sex determining region Y)-box 18	9.97	13.52	115.65	34.92
sox7	SRY (sex determining region Y)-box 7	479.79	386.91	730.47	309.74
spi1b	Spi-1 proto-oncogene b	478.85	65.53	201.04	31.29
spry2	sprouty RTK signaling antagonist 2	8.48	21.10	33.97	25.18
spry4	sprouty homolog 4 (Drosophila)	51.40	65.37	73.22	42.41
supt6h	SPT6 homolog, histone chaperone	22.42	22.80	23.26	19.04
yap1	Yes-associated protein 1	25.54	13.78	38.39	8.58

Table of haematopoietic and vascular transcription factors identified through protein class analysis of genes expressed at 2 FPKM or greater, in all four early *scf*<sup>+</sup> populations. Biological GO term analysis was carried out on expressed-in-all genes, to identify factors associated with haematopoietic and vascular processes. Protein class analysis was then carried out on the resulting gene list, to isolate a expressed-in-all total transcription factor list.

Tissue-specific biological processes relating to the signalling molecule class of commonly expressed proteins include negative regulation of axon extension, vasculogenesis and angiogenesis.

**Figure 4-14 Scatter plot of genes with the same expression level across all four early *sc<sup>l+</sup>* contexts.**



I have also generated a list of genes that are similarly expressed in each of the *sc<sup>l+</sup>* populations, these are a subset of the previous analysis- not only are these genes expressed (FPKM>2) in all four contexts but at a similar level of expression in all four contexts. Genes were included in the list if they showed non-significant differential expression following DESeq2 analysis for the anterior, posterior and 10ss comparisons, plus showed expression levels of FPKM>2 in all contexts.

5344 genes met these criteria and represented 204 biological GO terms. Figure 4-14 shows that the majority of these genes show expression under 1000 FPKM in each context. Only one major biological tissue-specific term was over-represented (other over-represented biological GO terms related to ubiquitous housekeeping processes or general development). 78 genes (detailed in Table 4-4) with a mean expression level of over 150 FPKM were associated with the over-represented tissue-specific process, haematopoiesis.

**Table 4-4 Haematopoietic genes showing common expression levels across the four early *scl* expressing populations studied.**

Gene Name	Description	10ss Ant. Mean FPKM	10ss Post. Mean FPKM	10ss Ant. Mean FPKM	20ss Post. Mean FPKM
rp119	ribosomal protein L19	1981.84	2246.24	2440.86	3482.23
etv2	ets variant 2	1535.57	788.51	1384.75	598.63
fosab	v-fos FBJ murine osteosarcoma viral oncogene homolog Ab	1437.93	1182.45	1003.04	873.22
rplp1	ribosomal protein, large, P1	1359.89	1461.90	1846.93	2349.08
rps7	ribosomal protein S7	1336.61	1358.00	1228.76	1623.16
rps27.1	ribosomal protein S27, isoform 1	1059.86	1106.44	1187.86	1484.69
rpl35	ribosomal protein L35	827.82	752.01	617.95	766.26
rps14	ribosomal protein S14	597.40	644.89	714.99	873.37
rpl11	ribosomal protein L11	453.58	480.25	524.88	596.85
rpl27	ribosomal protein L27	450.65	432.34	487.95	580.05
rps19	ribosomal protein S19	436.28	418.55	468.50	589.76
rpl22	ribosomal protein L22	406.21	347.77	392.72	475.81
rps29	ribosomal protein S29	341.17	362.30	390.02	483.80
hdac1	histone deacetylase 1	330.71	367.55	263.17	273.59
hsp70l	heat shock cognate 70-kd protein, like	324.03	206.62	184.03	117.65
ube2b	ubiquitin-conjugating enzyme E2b	247.14	226.54	174.08	188.50
rpl35a	ribosomal protein L35a	175.60	196.02	255.02	313.08
wdr43	WD repeat domain 43	165.67	234.17	108.07	236.01
atp11a	ATPase inhibitory factor 1a	127.84	111.20	127.00	145.26
sac1	SUMO1 activating enzyme subunit 1	117.80	115.65	69.23	75.45
rps24	ribosomal protein S24	95.25	93.69	85.81	105.76
ddx18	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18	93.88	135.32	75.56	168.20
smad5	SMAD family member 5	83.31	103.78	127.93	75.49
ak2	adenylate kinase 2	67.74	107.61	106.56	139.38
snrnp70	small nuclear ribonucleoprotein 70 (U1)	63.40	55.77	43.64	33.93
rcor1	REST corepressor 1	61.36	39.38	49.73	27.48
cxcl12b	chemokine (C-X-C motif) ligand 12b (stromal cell-derived factor 1)	57.07	52.56	51.38	24.02
hspa9	heat shock protein 9	54.34	90.83	64.35	125.10
spry4	sprouty homolog 4 (Drosophila)	51.40	65.37	73.22	42.41
cdc73	cell division cycle 73, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)	45.78	50.13	37.91	45.92
ube2a	ubiquitin-conjugating enzyme E2a	44.38	42.39	30.61	38.04
brd2a	bromodomain containing 2a	44.21	60.02	43.45	45.31
brd3a	bromodomain containing 3a	42.50	55.05	38.18	39.03
ptenb	phosphatase and tensin homolog B	40.99	30.42	61.99	24.90
cbfb	core-binding factor, beta subunit	33.82	21.20	49.49	15.83
epsf1	cleavage and polyadenylation specific factor 1	29.27	40.69	36.57	47.09
jag1b	jagunal homolog 1b	28.00	31.37	26.37	33.56
taf3	TAF3 RNA polymerase II, TATA box binding protein (TBP)-associated facto	26.18	32.19	25.53	24.38
yap1	Yes-associated protein 1	25.54	13.78	38.39	8.58
sart3	squamous cell carcinoma antigen recognized by T cells 3	24.23	30.15	21.36	22.51
acvr1l	activin A receptor, type I like	23.71	24.02	20.40	17.32
sbds	Shwachman-Bodian-Diamond syndrome	22.72	24.11	16.08	23.43
kras	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog	22.71	24.55	19.15	15.52
nfil3	nuclear factor, interleukin 3 regulated	22.29	11.32	34.36	11.38
pdc2	programmed cell death 2	19.88	20.41	16.00	34.32
slc48a1b	solute carrier family 48 (heme transporter), member 1b	17.51	11.77	9.46	7.07
dhx8	DEAH (Asp-Glu-Ala-His) box polypeptide 8	17.48	26.31	20.40	24.36
sec23b	Sec23 homolog B, COPII coat complex component	17.06	17.33	25.04	18.85
melk	maternal embryonic leucine zipper kinase	16.27	13.93	8.59	15.42
aggf1	angiogenic factor with G patch and FHA domains 1	15.70	17.21	12.49	16.42
ncor2	nuclear receptor corepressor 2	14.93	11.65	22.29	11.77
ptena	phosphatase and tensin homolog A	12.36	11.67	20.08	14.28
gna13b	guanine nucleotide binding protein (G protein), alpha 13b	11.87	16.52	21.73	14.30
brd4	bromodomain containing 4	11.76	13.37	17.88	11.01
cxcl1b	CXCL1 chemokine	11.40	8.31	8.91	11.78
brf1b	BRF1, RNA polymerase III transcription initiation factor b	11.12	8.15	9.35	9.20
pttg1ipb	pituitary tumor-transforming 1 interacting protein b	9.48	6.61	9.64	4.11
adnp2b	ADNP homeobox 2b	9.25	11.33	14.96	11.53
tert	telomerase reverse transcriptase	8.44	8.02	11.82	10.27
sfxn4	sideroflexin 4	8.44	5.55	8.30	7.11
pak4	p21 protein (Cdc42/Rac)-activated kinase 4	7.91	12.53	8.26	11.21
adnp2a	ADNP homeobox 2a	7.68	8.04	8.46	6.47
mlh1	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	7.63	8.35	6.04	7.74
wasla	Wiskott-Aldrich syndrome-like a	7.37	8.39	5.50	3.14
adnpb	activity-dependent neuroprotector homeobox b	6.59	5.30	5.48	4.24
pdzk1ip1	PDZK1 interacting protein 1	6.45	4.89	3.80	3.14
vhl	von Hippel-Lindau tumor suppressor	6.42	6.57	11.86	8.96
clpxa	caseinolytic mitochondrial matrix peptidase chaperone subunit a	6.35	8.29	10.64	7.59
tet3	tet methylcytosine dioxygenase 3	6.32	4.04	12.74	5.01
cas3a	caspace 3, apoptosis-related cysteine peptidase a	6.26	6.66	6.35	5.90
abcc5	ATP-binding cassette, sub-family C (CFTR/MRP), member 5	5.64	4.63	10.38	4.53
fev	FEV (ETS oncogene family)	4.42	5.18	3.73	3.40
slc25a38b	solute carrier family 25, member 38b	4.33	5.00	3.11	8.38
waslb	Wiskott-Aldrich syndrome-like b	3.76	5.26	4.21	3.84
mfn2	mitofusin 2	3.69	5.88	3.98	5.37
brf1a	BRF1, RNA polymerase III transcription initiation factor a	3.31	3.71	3.90	2.02
bmp4	bone morphogenetic protein 4	3.18	3.15	2.29	5.55
sh2b3	SH2B adaptor protein 3	2.73	3.98	3.85	2.84

Table of haematopoietic genes identified as commonly expressed in all four early *scl*<sup>+</sup> populations. List is in order of expression level in the 10ss anterior.

The genes shown in table 4-4 include several factors that show ubiquitous activity, such as the genes for ribosomal proteins and *hdac1*. These factors are erroneously

included due to inaccurate annotation in the zfin database- for instance *hdac1* is annotated with 17 different tissue-specific functions, only one of which is haematopoietic. This example highlights the major issue with the gene ontology enrichment analysis, however this approach remains the most appropriate to summarise genome-wide transcriptomic variation.

9.8% of these commonly expressed factors were classed as transcription factors by PANTHER database and 2.5% as signalling molecules. The biological processes relating to these transcription factors were for ubiquitous processes including regulation of the cell-cycle kinases, chromatin remodelling and steroid signalling. The signalling molecule class showed enriched biological GO terms for apoptotic processes and negative regulation of Ras signalling.

#### 4.5. Enriched *scf*<sup>+</sup> expression profiles compared to *scf* neighbouring cells.

The previous comparative transcriptomic analysis of these *scf*-expressing populations showed some differentially expressed genes were associated with haematopoietic biological GO categories, however a large number of general developmental genes accounted for many of the differentially expressed factors. These semi-ubiquitous and often highly expressed genes made investigating *scf* cell specific transcriptomic features challenging as their great number swamped any analysis.

In zebrafish primitive erythropoiesis occurs primarily in the posterior of the embryo, thus it was expected that this *scf*<sup>+</sup> population at the 20ss would be strongly erythrogenic. Gene ontology analysis of genes classed as expressed in this context identified 279 significantly over-represented biological terms. Only 12 of these 279 biological terms relate to haematopoietic or blood vessel development. This demonstrates how ubiquitous processes can swamp enrichment analyses to a degree that impairs resolution of true biological function.

Here I describe my findings when investigating the features of enriched  $scf^+$  transcriptomes compared to context appropriate  $scf^-$  transcriptomes. Genes that are significantly enriched in  $scf^+$  cells can be identified and their biological function investigated. Historically it has been necessary to take a one gene at a time approach, and has made significant advances in our understanding of the emergence of the haematopoietic and vascular system. With the advent of genome-wide approaches such as those employed in this project I have begun to tackle this problem in a true *in vivo* setting on a systems basis. Thus the enriched genes lists are likely to not only include previously studied haemangiogenic factors but also novel proteins that may play key roles in this process.

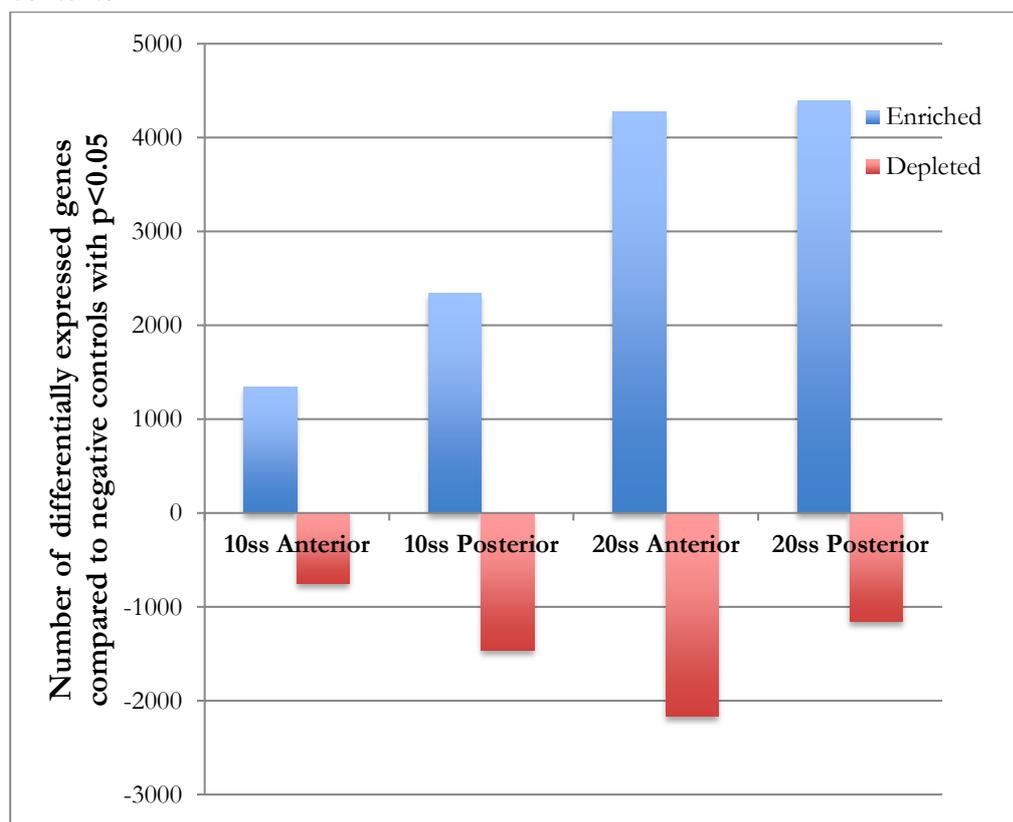
In this analysis I am focussing on the factors that are enriched in  $scf^+$  cells, but not on the depleted genes. While the enriched factors are specifically enriched in  $scf^+$  cells versus all other developing cell types within the anterior or posterior of the fish, the same logic cannot be applied to depleted genes. Though this differential expression analysis, genes that are specifically depleted in  $scf^+$  cells will not be resolved from genes that are specifically expressed in another cell type from the same context. Thus although  $scf$  expression has previously been shown to lead to down-regulation of the expression of specific targets, and it is known that the haemangioblast and Scl-core-complexes can represses other cell fates<sup>102,138,266</sup>, depleted factors identified in this analysis would be difficult to interpret.

#### 4.5.1. Differential expression between $scf^+$ and neighbouring $scf^-$ contexts

DESeq2 was used to assess the significance of difference in expression of genes between citrine positive ( $scf^+$  cells) and citrine negative ( $scf^-$  cells) cells, taking into account biological variation. Genes were classed as differentially expressed if the fold change between positive and negative samples for each context was greater than

expression differences between replicates, with a confidence of  $p < 0.05$ . Factors enriched in  $scl^+$  samples were also subjected to an expression cut-off of FPKM  $> 2$  in the positive samples, in order to focus the analysis on genes that are definitely expressed within the cells of interest. The FPKM  $> 2$  cut-off in  $scl^+$  samples was also applied to the depleted gene lists to refine the list down to factors that were expressed in  $scl^+$  cells but at a significantly lower level than in surrounding cells. Without this cut-off genes that are not expressed at this developmental stage, would associate with very large fold change values due to minute differences in background read levels, and swamp further analysis.

**Figure 4-15 Distribution of differentially expressed genes over the four early  $scl^+$  contexts**



As the  $scl^+$  population develops from 10ss to the 20ss the number of differentially expressed genes increases, indicating that each population is becoming more differentiated from its neighbouring cell types and expressing more cell specific factors instead of a limited number of general developmental genes (figure 4-15).

The 10ss anterior *sc<sup>l</sup><sup>+</sup>* population has the least number of significantly differentially expressed genes. An interpretation of this would be that this population is, on a transcriptome level, the closest to surrounding tissues and thus may represent a developmentally earlier stage of the *sc<sup>l</sup><sup>+</sup>* population. The 10ss posterior *sc<sup>l</sup><sup>+</sup>* population is known to arise earlier in development and displays almost double the number of differentially expressed genes, supporting this theory.

#### 4.5.2. 10ss Anterior

Upon carrying out DESeq2 analysis of the 10ss *sc<sup>l</sup><sup>+</sup>* anterior samples compared to the 10ss anterior citrine negative cell samples 2096 genes were identified as being differentially expressed.

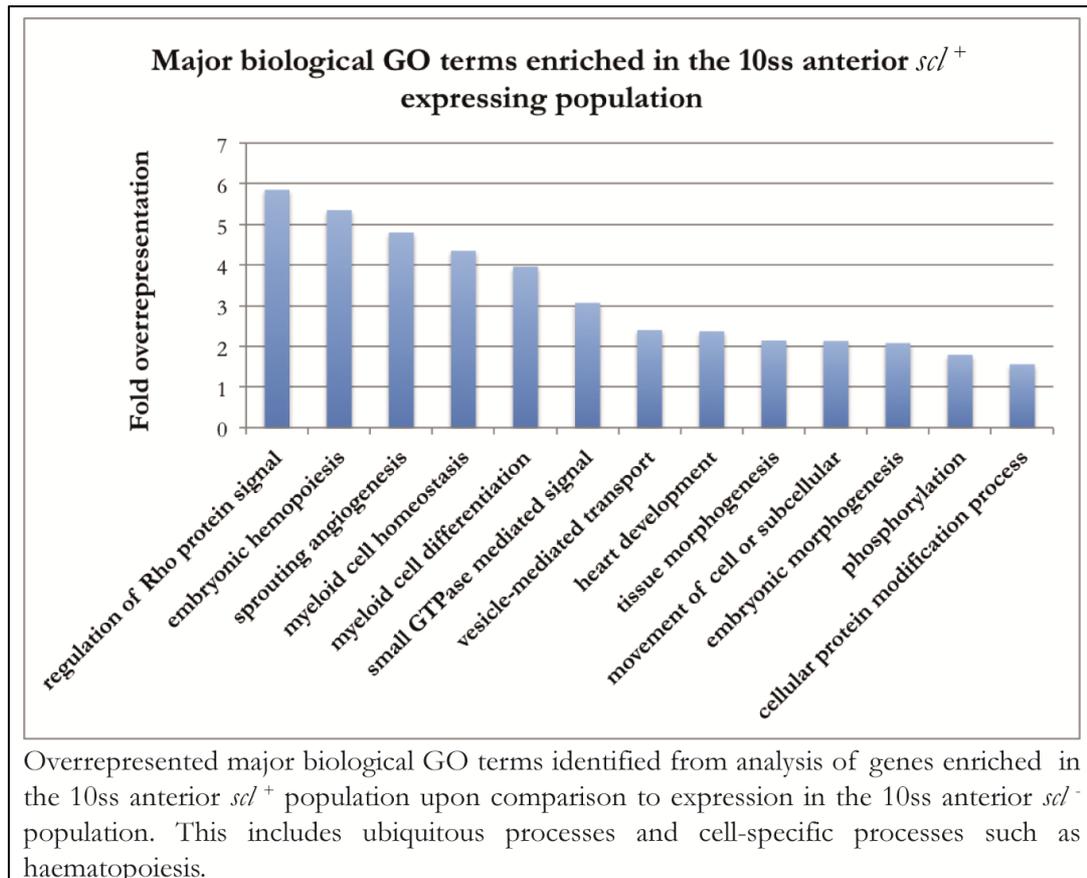
##### **10ss anterior enriched genes over negative controls**

Biological GO term analysis was carried out using the PANTHER database. 1346 genes showed enrichment within our *sc<sup>l</sup><sup>+</sup>* samples from the 10ss anterior of the zebrafish embryo.

This list was annotated with 56 different biological GO terms and included 291 unclassified genes enriched in this dataset. The term “unclassified” is used to label any gene that lacks a biological GO term classification, thus this category will include genes that have been identified but their biological function is unknown, novel genes and previously studied genes that lack representation in the GO database. As a result “unclassified” genes lists are an excellent starting point to investigate novel factors however such lists must be thoroughly checked against the literature.

11 over-represented biological GO terms related to tissue-specific processes, all of which were haematopoietic or vascular functions. The top over-represented biological GO terms are shown in Figure 4-16. Other enriched GO terms were for Rho signalling, small GTPase mediated signalling and vesicle mediated transport.

Figure 4-16 Over-represented tissue-specific gene ontology terms for genes enriched in the 10ss anterior *scl* expressing population.



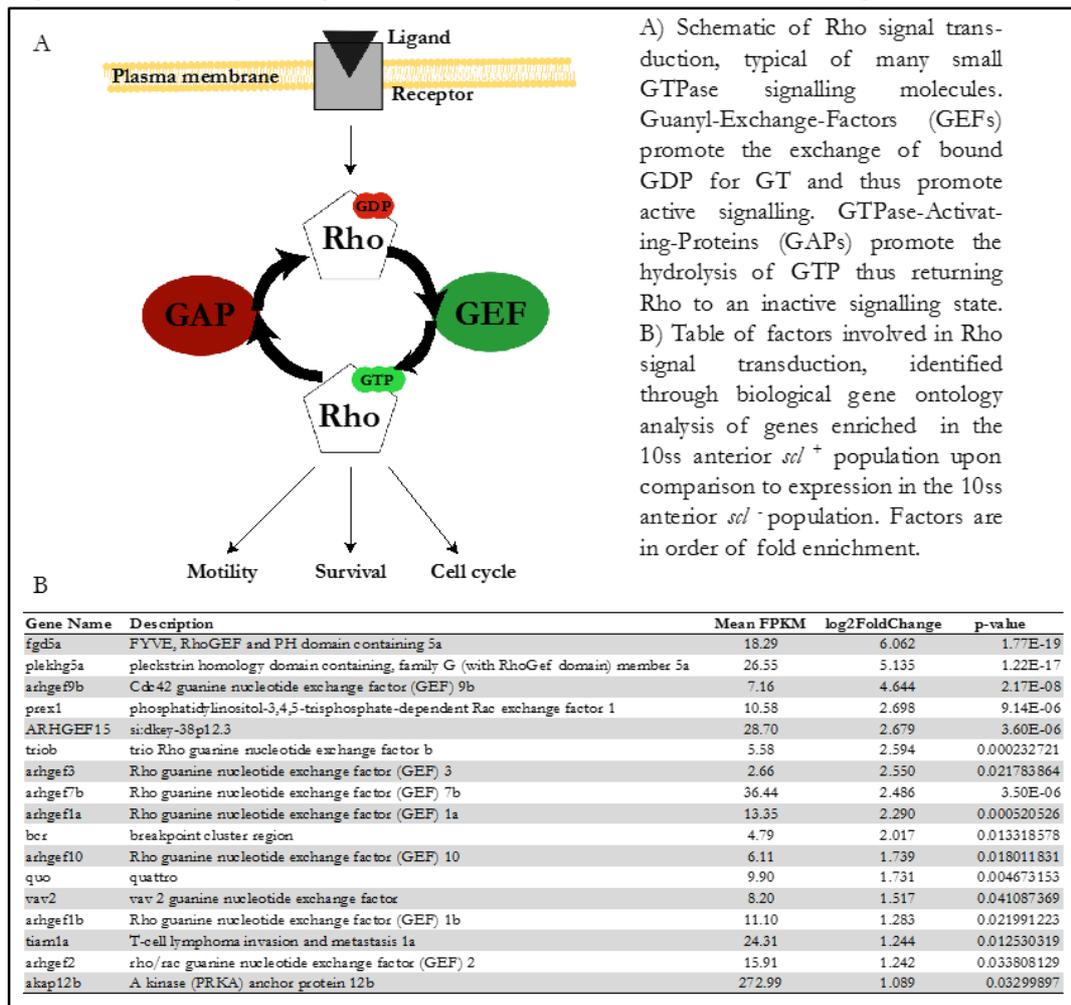
Rho signalling has been previously shown to contribute to proliferation and motility of multiple haematopoietic lineages<sup>267,268</sup>. Together this suggests that many of the genes specifically enriched in *scl*<sup>+</sup> cells of the anterior population contribute to biological processes related to a haemangiogenic profile.

### Rho signalling

Of the 17 genes annotated as being directly involved in Rho signal transduction, 11 show guanidine nucleotide exchange factor (GEF) activity. Figure 4-17 shows the identity of these Rho signalling proteins and their expression and enrichment level in the 10ss anterior *scl*<sup>+</sup> expressing population. A schematic of Rho signalling is also included indicating how the enrichment of GEFs would promote activation of Rho signalling cascades. Rho signalling has previously been shown to play crucial roles in regulating hematopoietic stem and progenitor cell (HSC/P) motility, survival and

proliferation. The haematopoietic roles of Rho GTPases Rac1 and Rac2 have been studied through the generation of knockout mice lines. Rac1 has been shown to be involved in HSC/P migration into the bone marrow<sup>269</sup> and following knockdown, the number of circulating HSC/P in mice embryos at E10.5, is dramatically reduced<sup>270</sup>. HSC/Ps from Rac2 deficient mice display a proapoptotic phenotype<sup>269</sup> plus Yang *et al.* have identified a role for this GTPase in regulating HSC/P adhesion<sup>271</sup>. The expression of these factors specifically within *scf*<sup>+</sup> cells at the 10ss indicates that Rho signalling may be important for the cellular identity of this population and may contribute to earlier stages of haematopoietic development than previously studied.

**Figure 4-17 Rho signalling enriched in the 10ss anterior *scf* expressing population.**



## Embryonic haematopoiesis

The second most over-represented biological GO term for factors enriched in *scl*<sup>+</sup> cells of the 10ss anterior, was embryonic haematopoiesis, with 12 genes contributing directly to this. Table 4-5 shows the expression and enrichment of these factors, which includes *lmo2* and *gata1a*, whose protein products have been proposed to form a key multi-protein DNA binding complex with Scl to mediate haemangiogenic gene regulation<sup>120,272,273</sup>. Many of these factors have been previously shown to have complex roles that can vary dramatically between different contexts. As a result it is difficult to draw any specific lineage conclusions from their enrichment, except that their presence strongly supports that this population has haematopoietic and vascular potential.

**Table 4-5 Embryonic haematopoietic genes enriched in the 10ss anterior *scl* expressing population.**

Gene Name	Description	Mean FPKM	log2FoldChange	p-value
<i>lmo2</i>	LIM domain only 2 (rhombotin-like 1)	3278.00	6.080	1.03E-23
<i>tal1</i>	T-cell acute lymphocytic leukemia 1	537.22	5.168	1.2801E-19
<i>etv2</i>	ets variant 2	1535.57	4.781	7.48E-24
<i>alas2</i>	aminolevulinate, delta-, synthase 2	10.79	3.749	1.35E-06
<i>sptb</i>	spectrin, beta, erythrocytic	13.17	3.658	8.07E-08
<i>gfi1aa</i>	growth factor independent 1A transcription repressor a	148.74	3.174	1.77E-09
<i>tbx20</i>	T-box 20	21.48	3.034	3.42E-05
<i>tmem88a</i>	transmembrane protein 88 a	87.84	2.655	4.01E-05
<i>erg</i>	v-ets avian erythroblastosis virus E26 oncogene homolog	10.69	2.612	0.00118299
<i>cebpa</i>	CCAAT/enhancer binding protein (C/EBP), alpha	107.65	2.452	3.35E-05
<i>myct1a</i>	myc target 1a	4.91	2.450	0.02290766
<i>caf1</i>	ELL associated factor 1	33.31	2.163	0.00132767
<i>gata1a</i>	GATA binding protein 1a	50.29	2.088	0.00123165

Table of embryonic haematopoietic factors identified through biological gene ontology analysis of genes enriched in the 10ss anterior *scl*<sup>+</sup> population upon comparison to expression in the 10ss anterior *scl*<sup>-</sup> population. Factors are in order of fold enrichment.

ETV2 (a.k.a. *etsrp*) is a member of the ETS family of transcription factors, which play redundant roles in vasculature and myeloid development<sup>20,40-43</sup>. The *etv2* is the first ETS family member to be expressed within the zebrafish embryo and shows strong expression in the anterior lateral plate mesoderm even at 1ss<sup>20</sup>. *scl*, *spi1b* and *flk* expression in the ALM is lost following *etv2* morpholino knockdown<sup>20</sup> and results in severe myeloid deficiencies<sup>41</sup> and complete loss of circulation<sup>42-46</sup>.

Lmo2 forms heterodimers with the Scl protein<sup>56,125</sup> and acts in synergy with Scl to mediate blood vessel and erythroid development, and also to mediate severe T-ALL<sup>123,125,130</sup>.

Gfi1aa is a member of the highly conserved family of Snail zinc finger proteins that act to repress transcription of target genes through the recruitment of histone modifying proteins<sup>274</sup>. This transcription factor has been shown to promote the erythroid lineage and repress the myeloid fate, mediated through regulation of *gata1* and *spi1b* expression levels respectively<sup>275</sup>.

The gene *sptb* (a.k.a. riesling) was identified after mutation resulted in reduced blood cell counts in zebrafish embryos<sup>276</sup>. Liao *et al.* describe Sptb as a cytoskeletal protein that is expressed in erythroid cells and is required for terminal erythroid differentiation<sup>277</sup>.

#### Myeloid cell differentiation

The anterior of the zebrafish embryo has been known to be the main embryonic origin of the myeloid lineage<sup>278-280</sup>. This enrichment analysis of *scl*<sup>+</sup> cells within the 10ss anterior correlates with this conclusion as it identifies the myeloid cell differentiation GO term as over-represented by 3.37 fold. 20 genes are annotated as contributing directly to this process from this enriched gene list as detailed in table 4-6. Factors showing particularly high enrichment compared to neighbouring tissues include *lmo2*, *etv2*, *sptb*, *gfi1aa* in addition to *spi1b*, *gata5*, *alas2* and *gfi1ab*.

Spi1b (a.k.a. pu.1) is a key transcription factor regulating the myeloid and lymphoid lineages<sup>191,281-283</sup>, and has been shown to be capable of partial rescue of the myeloid lineage in the *cloche* mutant i.e. in the absence of *scl* expression<sup>205</sup>. Spi1b and GATA1 have been shown to cross-regulate each other forming a fate decision kernel within the regulatory network of haematopoietic development<sup>205</sup>. Expression of *spi1b* at this

magnitude strongly suggests that this *scf*<sup>+</sup> population has the potential to contribute to myeloid lineages.

**Table 4-6 Genes annotated with the GO term myeloid cell differentiation enriched in the 10ss anterior *scf* expressing population.**

Gene Name	Description	Mean FPKM	log2FoldChange	p-value
<i>spi1b</i>	Spi-1 proto-oncogene b	478.85	6.379	1.92E-34
<i>lmo2</i>	LIM domain only 2 (rhombotin-like 1)	3278.00	6.080	1.03E-23
<i>tal1</i>	T-cell acute lymphocytic leukemia 1	537.22	5.168	1.2801E-19
<i>etv2</i>	ets variant 2	1535.57	4.781	7.48E-24
<i>gata5</i>	GATA binding protein 5	146.72	4.308	3.71E-14
<i>gfi1ab</i>	growth factor independent 1A transcription repressor b	112.13	4.080	4.25E-08
<i>alas2</i>	aminolevulinate, delta-, synthase 2	10.79	3.749	1.35E-06
<i>sptb</i>	spectrin, beta, erythrocytic	13.17	3.658	8.07E-08
<i>gfi1aa</i>	growth factor independent 1A transcription repressor a	148.74	3.174	1.77E-09
<i>casp8</i>	caspase 8, apoptosis-related cysteine peptidase	12.59	2.973	0.00014378
<i>tmem88a</i>	transmembrane protein 88 a	87.84	2.655	4.01E-05
<i>slc25a38a</i>	solute carrier family 25, member 38a	6.57	2.586	0.01068619
<i>smad9</i>	SMAD family member 9	18.24	2.574	0.00015831
<i>irf8</i>	interferon regulatory factor 8	15.65	2.451	0.00247817
<i>gata1a</i>	GATA binding protein 1a	50.29	2.088	0.00123165
<i>rasa3</i>	RAS p21 protein activator 3	14.02	1.785	0.00536395
<i>sec23b</i>	Sec23 homolog B, COPII coat complex component	17.06	1.743	0.0029278
<i>slc25a38b</i>	solute carrier family 25, member 38b	4.33	1.633	0.04705176
<i>ncor1</i>	nuclear receptor corepressor 1	41.81	1.624	0.00334737
<i>kif1b</i>	kinesin family member 1B	4.75	1.391	0.0289401
<i>ptenb</i>	phosphatase and tensin homolog B	40.99	1.008	0.04857451

Table of myeloipoeitic factors identified through biological gene ontology analysis of genes enriched in the 10ss anterior *scf*<sup>+</sup> population upon comparison to expression in the 10ss anterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

GATA proteins are transcription factors that bind the DNA sequence (A/T)GATA(A/G). GATA factors 4, 5 and 6 contribute to the myeloid lineage and are essential for the formation of anterior haemangioblasts<sup>284</sup>. These factors have also been shown to contribute to cardiac and endothelial development<sup>285,286</sup> and migration of cardiac precursors into the developing heart field<sup>284,287,288</sup>. The *gata1* gene is observed to be expressed in developing erythroid cells, while *gata2* is expressed earlier in the lateral plate mesoderm and later in HSC and progenitors<sup>289</sup>. Enforced expression of *gata2* has been shown to partially compensate for a loss in GATA1 function, thus contributing to erythroid development<sup>82</sup>. Functional GATA2 is required for the expression of *runx1* within the endothelium<sup>290</sup> and thus is essential for endothelial-to-haematopoietic transition (EHT)<sup>224,291</sup>.

The *alas2* gene was identified through screening as the gene mutated in “sauternes” mutant zebrafish line, which displays anaemia as a result of defective haemoglobin synthesis<sup>292</sup>. The gene encodes an enzyme required for the initial step of haem biosynthesis and is specifically expressed in early erythroid cells<sup>293</sup>. The published data suggests that this is in fact a key erythroid gene rather than contributing to a myeloid fate as GO analysis has annotated it, in the PANTHER database. This highlights the dependency of GO enrichment analysis on the accuracy of the annotations. It is also interesting to note that key genes for both myeloid (*spi1b*) and erythroid (*alas2*) lineages are co-expressed and specifically enriched within a single early *scf*<sup>+</sup> population. This may represent a progenitor population co-expressing factors of opposing lineages or cellular heterogeneity within the population.

#### Angiogenesis

Angiogenic terms were also identified as an over-represented biological GO process, with 25 genes within this 10ss anterior *scf*<sup>+</sup> gene list contributing directly to this term (table 4-7). The over-representation of angiogenic terms rather than vasculogenic terms could be due to the considerable overlap of genes contributing to these two processes and poor annotation. Angiogenesis requires pre-existing vasculature from which new vessels can be formed. If angiogenesis is truly enriched within the 10ss anterior *scf*<sup>+</sup> population vessels must have formed by this stage- however no vessels are visible within the embryo at this stage. Conversely the majority of the factors included in this list contribute to a range of vascular developmental roles, and for some such as *etn2* also contribute to haematopoietic development.

Cdh5 (a.k.a. VE-cadherin) is a cell surface molecule expressed through vasculogenic, angiogenic angioblasts and endocardium development<sup>294</sup> that controls cell-cell adhesion<sup>295</sup>. Expression of *cdh5* is commonly used as a marker of endothelial cells<sup>11,296-298</sup>.

Plekhhg5a (a.k.a. syx-a) is a Rho A GEF required in zebrafish and mice for angiogenic sprouting<sup>299</sup>, and has been shown to be expressed throughout the vascular system. Rho signalling through Plkhg5a is necessary for the formation of the intersegmental vessels of the zebrafish tail<sup>300</sup>.

Fli1 is a member of the ETS transcription factor family and an early marker of vasculogenesis<sup>44</sup> initially being expressed in posterior haemangioblasts before becoming restricted to endothelial cells<sup>301</sup>. Loss of function studies have shown that *fli1* cooperates with *erg* to potentially promote Cdh5 activity<sup>20</sup>. Erg is also enriched in the 10ss anterior *scf*<sup>+</sup> cells- this is another member of the ETS family and expressed blood, endothelial and pharyngeal arches<sup>45,302</sup>. Both have been shown to be required for angiogenesis, with morpholino-mediated knockdown resulting in haemorrhage in the head in both zebrafish and mice<sup>20,303</sup>. Erg has been shown to directly bind to the *cdh5* promoter in human cells, to promote expression in endothelial cells<sup>304</sup>.

**Table 4-7 Genes annotated with the GO term angiogenesis enriched in the 10ss anterior *scf* expressing population**

Gene Name	Description	Mean FPKM	log2FoldChange	p-value
cdh5	cadherin 5	48.95	7.104	4.30E-29
kdr	kinase insert domain receptor (a type III receptor tyrosine kinase)	23.14	6.193	4.27E-19
tal1	T-cell acute lymphocytic leukemia 1	537.22	5.168	1.28E-19
plekhhg5a	pleckstrin homology domain containing, family G member 5a	26.55	5.135	1.22E-17
etv2	ets variant 2	1535.57	4.781	7.48E-24
arhgef9b	Cdc42 guanine nucleotide exchange factor (GEF) 9b	7.16	4.644	2.17E-08
hspa12b	heat shock protein 12B	5.63	4.092	1.06E-05
flt1	fms-related tyrosine kinase 1	6.41	4.071	2.57E-07
fli1a	Fli-1 proto-oncogene, ETS transcription factor a	151.39	4.022	1.12E-10
nrp1b	neuropilin 1b	35.84	3.673	1.94E-11
fn1a	fibronectin 1a	168.73	3.019	4.17E-09
ramp2	receptor (G protein-coupled) activity modifying protein 2	38.99	2.737	8.38E-06
dll4	delta-like 4 (Drosophila)	14.13	2.694	0.00017359
erg	v-ets avian erythroblastosis virus E26 oncogene homolog	10.69	2.612	0.00118299
arhgef7b	Rho guanine nucleotide exchange factor (GEF) 7b	36.44	2.486	3.50E-06
mcamb	melanoma cell adhesion molecule b	19.93	2.419	0.00039005
itga5	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	45.88	2.134	3.35E-05
cds2	CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2	41.99	2.065	0.00046752
sat1b	spermidine/spermine N1-acetyltransferase 1b	12.91	1.911	0.01933028
cdc42	cell division cycle 42	131.66	1.835	8.48E-05
itgav	integrin, alpha V	5.98	1.834	0.0168433
fimn3	formin-like 3	20.14	1.520	0.00475832
amotl2a	angiomin like 2a	48.00	1.489	0.00233987
hapln1b	hyaluronan and proteoglycan link protein 1b	23.10	1.456	0.00615165
shc1	SHC (Src homology 2 domain containing) transforming protein 1	15.80	1.294	0.02521639
unc5b	unc-5 netrin receptor B	11.78	1.231	0.02602951

Table of angiogenic factors identified through biological gene ontology analysis of genes enriched in the 10ss anterior *scf*<sup>+</sup> population upon comparison to expression in the 10ss anterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

In the 10ss anterior *scf*<sup>+</sup> population I have identified enrichment of *kdr* (*VEGFR-2*) and *flt1* (*VEGFR-1*) transcripts, compared to their expression in surrounding tissues. *kdr* is the first of the VEGFRs to be expressed in zebrafish and is expressed in the developing vasculature, in a pattern indistinguishable from key endothelial factor, *Kdr1* (*VEGFR-4*)<sup>33,305</sup>. Morpholino knockdown of *kdr* alone did not affect vasculogenesis or angiogenesis, but in combination with *kdr1* morpholino resulted in dramatic vascular defects suggesting a co-operative role for these two receptors in the proliferation and morphogenesis of embryonic vasculature<sup>305,306</sup>. The *flt1* (*VEGFR-1*) gene is expressed strongly in the developing vasculature and cardiomyocytes<sup>307</sup>. Knockdown of *flt1* results in vascular defects including dorsal aorta and posterior cardinal vein malformation<sup>308</sup> plus defects in arterial branching<sup>307</sup>. Loss of *flt1* also causes ventricular contractility to be lost by 72 hpf, indicating that this receptor is required for the maintenance of the heart beat<sup>308</sup>.

*Arhgef9b* is also a GEF, specifically acting on the Rho GTPase *Cdc42* to activate its signalling activity. Unlike the other factors included in this angiogenic gene list, *arhgef9b* has been specifically associated with angiogenesis, rather than a broader endothelial developmental role. Sprouting angiogenesis is stimulated via *arhgef9a* and *cdc42*, in response to BMP signals<sup>309</sup>.

### Conclusions

10ss anterior genes specifically enriched in *scf* expressing cells are annotated with a limited list of statistically over-represented biologically related GO terms. These notably include embryonic haematopoiesis, myelopoiesis and angiogenesis plus the Rho signalling pathway, which has previously been shown to be involved in the regulation of the aforementioned processes. The enriched factors include genes shown to be key regulators and signalling molecules in previous studies using *in situ* techniques, morpholino and over-expression assays in zebrafish, mice and cell

culture models. My RNA-seq analysis of this early *sc<sup>+</sup>* population has confirmed the co-expression of these factors within a single cell population, *in vivo* while also describing the full transcriptome of this population including previously unknown factors.

### Top enriched gene lists for the 10ss anterior, compared to negative controls

In the top 50 enriched genes in the 10ss anterior, 33 genes show some expression in the *sc<sup>-</sup>* sample whereas 17 fail to have any control transcripts mapping to them thus can result in misleadingly high enrichment values. Of these 33, 17 genes (detailed in table 4-8) have been previously shown to contribute to haematopoietic or vascular development.

**Table 4-8 Top enriched genes in the 10ss anterior *sc<sup>+</sup>* expressing population.**

Gene Name	Description	Mean FPKM	log <sub>2</sub> FoldChange	p-value	Novel?
<i>cdh5</i>	cadherin 5	48.95	7.104	4.30E-29	
<i>map4k2</i>	mitogen-activated protein kinase kinase kinase kinase 2	14.06	6.934	3.10E-14	
<i>ncam3</i>	neural cell adhesion molecule 3	25.15	6.775	1.72E-15	NOVEL
<i>VAV3</i>	<i>sich73-383g2.1</i>	21.89	6.759	2.00E-16	
<i>hyal2a</i>	hyaluronoglucosaminidase 2a	43.92	6.707	2.93E-17	
<i>spi1b</i>	Spi-1 proto-oncogene b	478.85	6.379	1.92E-34	
<i>sidkey-37g12.1</i>	<i>sidkey-37g12.1</i>	6.44	6.343	7.64E-12	NOVEL
<i>kdr</i>	kinase insert domain receptor (a type III receptor tyrosine kinase)	23.14	6.193	4.27E-19	
<i>kdl1</i>	kinase insert domain receptor like	53.84	6.190	4.47E-22	
<i>calcr1a</i>	calcitonin receptor-like a	45.90	6.169	3.99E-18	
<i>lmo2</i>	LIM domain only 2 (rhombotin-like 1)	3278.00	6.080	1.03E-23	
<i>fgd5a</i>	FYVE, RhoGEF and PH domain containing 5a	18.29	6.062	1.77E-19	
<i>esama</i>	endothelial cell adhesion molecule a	81.39	5.982	1.24E-21	
<i>grapa</i>	GRB2-related adaptor protein a	31.72	5.828	8.32E-12	
<i>rpz4</i>	rapunzel 4	10.89	5.728	3.20E-09	NOVEL
<i>tpd52l1</i>	tumor protein D52-like 1	31.77	5.687	1.33E-15	
<i>fli1b</i>	Fli-1 proto-oncogene, ETS transcription factor b	65.15	5.512	6.59E-20	
<i>CSGALNACT1</i>	chondroitin sulfate N-acetylgalactosaminyltransferase 1	4.77	5.505	1.50E-09	
<i>tie1</i>	tyrosine kinase with immunoglobulin-like and EGF-like domains 1	41.59	5.486	2.78E-16	
<i>sox7</i>	SRY (sex determining region Y)-box 7	479.79	5.478	5.42E-28	
<i>egfl7</i>	EGF-like-domain, multiple 7	484.45	5.474	1.18E-26	
<i>tmem26a</i>	transmembrane protein 26a	29.55	5.473	3.68E-12	NOVEL
<i>cx45.6</i>	connexin 45.6	8.92	5.428	6.15E-08	
<i>rnd1</i>	Rho family GTPase 1	41.12	5.360	3.47E-18	
<i>gata4</i>	GATA binding protein 4	25.36	5.233	2.80E-10	
<i>robo2</i>	roundabout, axon guidance receptor, homolog 2 (Drosophila)	38.16	5.200	5.30E-08	
<i>tal1</i>	T-cell acute lymphocytic leukemia 1	537.22	5.168	1.28E-19	
<i>ikzf1</i>	IKAROS family zinc finger 1 (Ikaros)	13.70	5.158	2.18E-08	
<i>spns2</i>	spinster homolog 2 (Drosophila)	24.40	5.144	4.34E-14	
<i>plekhg5a</i>	pleckstrin homology domain containing, family G	26.55	5.135	1.22E-17	
<i>iqgap2</i>	IQ motif containing GTPase activating protein 2	19.54	4.928	8.85E-14	NOVEL
<i>klhl4</i>	kelch-like family member 4	56.98	4.899	1.73E-13	NOVEL
<i>lhx1b</i>	LIM homeobox 1b	23.87	4.876	5.48E-09	

Table of the top enriched genes in the 10ss anterior *sc<sup>+</sup>* population upon comparison to expression in the 10ss anterior *sc<sup>-</sup>* population. Factors are in order of fold enrichment. Novel genes have not been previously reported in the literature.

The remaining top enriched factors include *robo2*, an axon guidance receptor, plus 15 genes that have not been studied in relation to haematopoietic development, of

which 11 lack any functional testing. It is surprising that such strongly enriched factors in a key developmental population of a popular model organism remain to be investigated. This is especially exciting as 17 of the 22 studied genes in this list play important roles related to haemangiogenesis.

### **Molecular function and protein class analysis of 10ss anterior enriched genes over negative controls**

Molecular function analysis correlated with biological function analysis revealed that the top significantly over-represented molecular function for the genes enriched in the 10ss anterior *scf*<sup>+</sup> cells is an involvement in Rho signalling. This includes factors showing Guanyl-nucleotide exchange, GTPase activator and GTP binding functions. Motor activity was also a highly over-represented molecular function for this gene list, which could suggest that this *scf*<sup>+</sup> population maybe particularly motile in comparison to its neighbouring cell types in the 10ss anterior. High motility specifically of this population is visible in time-lapse studies (Figure 3-9).

### **Protein class analysis of 10ss anterior enriched genes over negative controls**

6.6 and 6.5% of genes were annotated as signalling molecules or transcription factors, respectively, when protein class analysis was carried out on the 10ss *scf*<sup>+</sup> anterior enriched gene list using the PANTHER database. Upon biological GO term analysis of the transcription factors, the top over-represented terms are similar to those of the entire gene list, including vasculogenesis and myeloid cell development.

The GO term vasculogenesis, which is the developmental predecessor of angiogenesis, was annotated with 5 key transcription factors specifically enriched in this *scf*<sup>+</sup> context and have been previously studied to great extents through their contribution to vasculogenesis and angiogenesis. These are the angiogenic genes

*sox18*, *sox7*, *tbx20* and *hey2*, plus the vasculogenic gene *etv2*. The genes *sox18* and *sox7* have been shown to be functionally redundant but together play a crucial role in establishing the correct arteriovenous identity of developing endothelial cells<sup>310</sup>. *tbx20* (previously known as *brT*) is expressed within the anterior lateral plate mesoderm in cells also expressing *gata4*<sup>311</sup>. Functional *tbx20* is required for correct cardiovascular development and disruption of this gene abrogates circulation in the developing zebrafish embryo<sup>312</sup>. *hey2* (a.k.a. *gridlock*) is required for full circulation and is expressed in the lateral plate mesoderm before the formation of blood vessels<sup>313</sup>. The Hey2 protein acts downstream of Notch signalling to regulate the arteriovenous fate decision, favouring the formation of arterial vessels<sup>314</sup>.

### 10ss anterior depleted genes compared to negative controls

Analysis of the PANTHER database determined 144 biological terms to be significantly depleted in the *scf*<sup>+</sup> cells compared to their *citrine*<sup>-</sup> expressing neighbours of the 10ss anterior. These biological functions related to both ubiquitous and tissue-specific processes. Ribosome biogenesis was dramatically depleted, as were other processes that contribute to protein translation and ATP synthesis. Interestingly regulation of DNA-templated transcription was also included in depleted biological processes. Upon analysis of the functions of these genes alone, the PANTHER database mostly annotated these as tissue-specific transcription factors contributing to brain and sensory organ development and thus contribute to both the ubiquitous GO term of regulation of transcription and towards tissue-specific activities. Depleted tissue-specific terms included diencephalon and hindbrain development, sensory organ morphogenesis and eye-development. This correlates with the current knowledge of the role of *scf* expressing cells not contributing to the development of these tissues, at this early stage. From my earlier comparative analysis of the 10ss *scf*<sup>+</sup> anterior versus posterior transcriptomes,

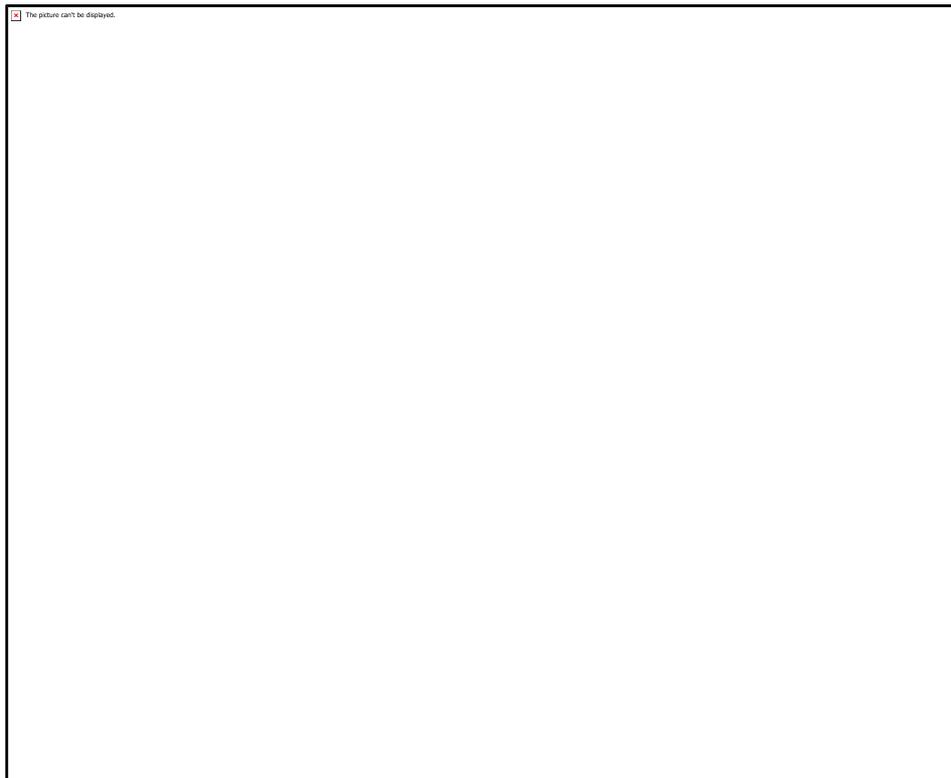
anterior enriched gene list included brain and eye development factors, this 10ss anterior *scf*<sup>+</sup> enrichment analysis revealed that this is residual expression that is significantly lower than in surrounding tissues.

#### 4.5.3. 10ss Posterior

DESeq2 analysis of the citrine<sup>+</sup> (*scf* expressing) samples, and the citrine negative samples, for the 10ss posterior identified 3805 genes showing significantly different expression between the *scf*<sup>+</sup> population and its posterior neighbouring cells.

As previously, annotation from the PANTHER database was used to assess the over-representation of biological GO terms associated with factors contained within enriched and depleted gene lists.

**Figure 4-18 Over-represented tissue-specific gene ontology terms for genes enriched in the 10ss posterior *scf* expressing population.**



## 10ss posterior enriched genes over negative controls

2342 genes showed significant enrichment in the 10ss *scf*<sup>+</sup> cells of the posterior compared to surrounding non-*scf* expressing cells. These enriched genes related to 77 over-represented biological GO categories and included 536 unclassified genes- the most significantly over-represented biological GO terms are detailed in figure 4-18. 14 over-represented biological GO terms related to tissue-specific processes- all were haematopoietic or vascular functions.

### Oxygen transport

9 globin genes plus myoglobin were significantly enriched in the *scf*<sup>+</sup> cells of the 10ss posterior. This included the subunits for embryonic haemoglobin alpha, beta and zeta proteins. Haemoglobin zeta, a form of haemoglobin alpha that is, in humans, produced in the yolk sac early in embryogenesis<sup>315</sup>. The enriched expression of these haemoglobin genes (listed in table 4-9) indicates that even at the 10ss the posterior *scf*<sup>+</sup> population has not only acquired an erythroid profile but also differentiated down this lineage to a point where functional oxygen transporter factors are produced at significant levels.

**Table 4-9 Genes annotated with the GO term oxygen transport enriched in the 10ss posterior *scf* expressing population**

Gene Name	Description	Mean fpkm	log2FoldChange	p-value
HBZ	hemoglobin zeta	12.24	7.0112	3.43E-12
hbae0	hemoglobin, alpha embryonic 1	132.99	4.9951	3.94E-27
hbbe2	hemoglobin beta embryonic-2	43.52	4.6407	6.17E-14
hbbe1.1	hemoglobin beta embryonic-1.1	216.93	4.4097	8.73E-29
hbbe3	hemoglobin beta embryonic-3	348.91	4.2972	6.66E-30
hbz	hemoglobin zeta	8.70	4.2817	4.30E-10
hbae1	hemoglobin, alpha embryonic 1	7.54	3.8685	1.79E-06
hbae3	hemoglobin alpha embryonic-3	407.10	3.8285	6.29E-34
hbbe1.2	hemoglobin beta embryonic-1.2	5.89	3.7114	9.87E-05
hbbe1.1	hemoglobin beta embryonic-1.1	44.15	2.7979	2.27E-08
mb	myoglobin	7.97	1.9136	0.000945036
hbae1	hemoglobin, alpha embryonic 1	19.47	1.8940	0.000246411

Table of factors involved in oxygen transport, identified through biological gene ontology analysis of genes enriched in the 10ss posterior *scf*<sup>+</sup> population upon comparison to expression in the 10ss posterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

## Erythroid development

In addition to the expression of the haemoglobin genes, the biological GO term erythroid development was also over-represented in the 10ss posterior *scf* expressing cells. This could suggest that the erythroid compartment at the 10ss is continuing to expand with this *scf* expressing population possibly containing erythroid progenitors at different stages of maturation. Genes contributing to this GO term are detailed in Table 4-10 and includes *gata1a*, *lmo2* and *sptb* which were enriched over 50 fold in the posterior *scf*<sup>+</sup> cells at the 10ss. Other enriched genes included *slc25a37* and *fech*.

**Table 4-10 Genes annotated with the GO term erythroid development enriched in the 10ss posterior *scf* expressing population.**

Gene Name	Description	Mean fpkm	log2FoldChange	p-value
<i>gata1a</i>	GATA binding protein 1a	634.37	7.610	1.84E-116
<i>tal1</i>	T-cell acute lymphocytic leukemia 1	627.24	6.699	3.87E-45
<i>lmo2</i>	LIM domain only 2 (rhombotin-like 1)	3763.49	6.369	1.22E-52
<i>sptb</i>	spectrin, beta, erythrocytic	17.79	6.095	4.16E-08
<i>spi1b</i>	Spi-1 proto-oncogene b	65.53	5.572	2.48E-40
<i>alas2</i>	aminolevulinate, delta-, synthase 2	34.11	5.233	2.16E-18
<i>slc25a37</i>	solute carrier family 25, member 37	15.25	3.151	2.15E-09
<i>fech</i>	ferrochelatase	25.91	2.749	7.38E-07
<i>tmod2</i>	tropomodulin 2	38.27	2.555	1.66E-06
<i>sec23b</i>	Sec23 homolog B, COPII coat complex component	17.33	1.324	0.00061073
<i>gata2a</i>	GATA binding protein 2a	25.33	1.024	0.0019461
<i>kmt2a</i>	lysine (K)-specific methyltransferase 2A	8.16	0.782	0.03976353

Table of factors involved in erythroid development, identified through biological gene ontology analysis of genes enriched in the 10ss posterior *scf*<sup>+</sup> population upon comparison to expression in the 10ss posterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

*Slc25a37* is a mitochondrial iron transporter that was initially identified in zebrafish when its gene became mutated caused blood cells to be lost during the first 4 days of development<sup>276</sup>. This transporter was suggested to be involved in maintaining the proliferation or avoidance of apoptosis during erythroid development as an absence of *slc25a37* causes a block at the proerythroblast stage<sup>316</sup>.

*fech* (a.k.a. *dracula*) was identified as a gene that upon mutation caused erythroid cells to be highly photosensitive and lyse upon normal exposure to light<sup>317</sup>. The *fech* gene encodes ferrochelatase, the final enzyme in the haem biosynthetic pathway, and is expressed in the posterior lateral plate mesoderm eventually becoming restricted to circulating erythrocytes<sup>318</sup>.

### Ras protein signalling

The Ras family of signalling molecules are central to growth regulatory signalling both within and outside of haematopoietic development and act downstream of tyrosine kinases such as the VEGFR family. Active Ras stimulates Raf proteins and MAPK kinase cascades, which transduce the signal into the cytoplasm and nucleus of the cell from Ras' location at the cell membrane. Key haematopoietic targets of MAPK cascade signalling include the GATA family proteins- both GATA1 and 2 are phosphorylated by MAPK signalling<sup>319,320</sup>. One of the enriched factors that contribute to Ras signalling in the 10ss *scf*<sup>+</sup> posterior population is *vav*, which is downstream of Ras, and can stimulate Rho signal transduction. In humans *vav* expression is restricted to haematopoietic cells<sup>321</sup> and there is evidence that Vav proteins are required for myeloid maturation<sup>322</sup>.

### Blood vessel morphogenesis

Unlike in the 10ss anterior *scf* expressing population where angiogenesis and vasculogenesis were over-represented terms, the 10ss posterior, enriched factors include genes annotated with the GO term blood vessel morphogenesis. This difference is difficult to interpret as many factors involved in vascular development, as already noted, are inaccurately annotated and are arbitrarily associated with a range of vascular processes. If blood vessel morphogenesis truly is particularly active in this context it would suggest that the posterior blood vessels are undergoing significant physical rearrangement. This correlates with the *scf* expression in the complete trunk vascular system as it forms (Figures 3- 7, 8 and 9) and with previously published data<sup>323</sup>. Genes contributing to this biological process are listed in table 4-11 and include the key genes *fli1(a & b)*, *kdr* and *sox7* plus *ecscr*, *npr1b* and *hspa12b*.

**Table 4-11 Genes annotated as involved in blood vessel morphogenesis enriched in the 10ss posterior *scl* expressing population.**

Gene Name	Description	Mean fpkm	log2FoldChang	p-value
<i>ecscr</i>	endothelial cell surface expressed chemotaxis and apoptosis regula	6.48	8.553	1.00E-20
<i>kdr</i>	kinase insert domain receptor (a type III receptor tyrosine kinase)	16.80	7.017	3.95E-41
<i>fli1b</i>	Fli-1 proto-oncogene, ETS transcription factor b	43.94	6.896	4.47E-68
<i>lmo2</i>	LIM domain only 2 (rhombotin-like 1)	3763.49	6.369	1.22E-52
<i>fli1a</i>	Fli-1 proto-oncogene, ETS transcription factor a	304.34	6.321	1.02E-66
<i>sox7</i>	SRY (sex determining region Y)-box 7	386.91	6.189	1.36E-64
<i>nrp1b</i>	neuropilin 1b	38.29	5.870	9.66E-61
<i>cdh5</i>	cadherin 5	19.64	5.691	2.41E-20
<i>plekhg5a</i>	pleckstrin homology domain containing, family G member 5a	12.20	5.489	1.39E-30
<i>arhgef9b</i>	Cdc42 guanine nucleotide exchange factor (GEF) 9b	9.76	5.318	1.25E-13
<i>flt1</i>	fms-related tyrosine kinase 1	3.38	5.258	2.01E-16
<i>etv2</i>	ets variant 2	788.51	5.081	1.51E-41
<i>hspa12b</i>	heat shock protein 12B	3.81	4.833	1.14E-09
<i>sox18</i>	SRY (sex determining region Y)-box 18	13.52	4.181	3.60E-15
<i>mcamb</i>	melanoma cell adhesion molecule b	49.77	3.675	3.09E-17
<i>colec12</i>	collectin sub-family member 12	43.73	2.823	1.52E-15
<i>ramp2</i>	receptor (G protein-coupled) activity modifying protein 2	45.70	2.624	3.01E-13
<i>lama4</i>	laminin, alpha 4	6.42	2.224	3.50E-07
<i>hey2</i>	hes-related family bHLH transcription factor with YRPW motif 2	78.62	2.137	6.08E-08
<i>dll4</i>	delta-like 4 (Drosophila)	9.22	1.977	0.001626
<i>snx5</i>	sorting nexin 5	56.42	1.943	6.45E-07
<i>rab13</i>	RAB13, member RAS oncogene family	48.66	1.920	1.14E-06
<i>mb</i>	myoglobin	7.97	1.914	0.000945
<i>itga5</i>	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	81.27	1.860	6.02E-07
<i>fmnl3</i>	formin-like 3	13.90	1.819	3.42E-08
<i>acvr1l</i>	activin A receptor type II-like 1	3.20	1.775	0.017373
<i>arhgef7b</i>	Rho guanine nucleotide exchange factor (GEF) 7b	36.30	1.640	1.62E-07
<i>gdf6a</i>	growth differentiation factor 6a	24.27	1.625	1.22E-05
<i>pld1a</i>	phospholipase D1a	2.85	1.464	0.021071
<i>gng2</i>	guanine nucleotide binding protein (G protein), gamma 2	36.64	1.431	0.000219
<i>itgav</i>	integrin, alpha V	6.51	1.345	0.002185
<i>ell</i>	elongation factor RNA polymerase II	14.56	1.342	0.000325
<i>cds2</i>	CDP-diacylglycerol synthase (phosphatidate cytidyltransferase) 2	34.26	1.306	0.000328
<i>shc1</i>	SHC (Src homology 2 domain containing) transforming protein 1	18.55	1.246	0.000231
<i>amotl2a</i>	angiomin like 2a	56.03	1.242	0.000374
<i>bmpr2a</i>	bone morphogenetic protein receptor, type II a (serine/threonine	2.09	1.197	0.033565
<i>amot</i>	angiomin	2.22	1.098	0.048437
<i>gata2a</i>	GATA binding protein 2a	25.33	1.024	0.001946
<i>msna</i>	moesin a	154.76	1.013	0.000424
<i>pik3c2a</i>	phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2	5.46	0.994	0.04574
<i>rbpj</i>	recombination signal binding protein for immunoglobulin kappa J	12.75	0.780	0.026701
<i>rab11a</i>	RAB11a, member RAS oncogene family	42.33	0.769	0.014109
<i>cdc42</i>	cell division cycle 42	111.96	0.763	0.009771
<i>kif11</i>	kinesin family member 11	53.63	0.750	0.048651

Table of factors involved in blood vessel morphogenesis, identified through biological gene ontology analysis of genes enriched in the 10ss posterior *scl*<sup>+</sup> population upon comparison to expression in the 10ss posterior *scl*<sup>-</sup> population. Factors are in order of fold enrichment.

*Ecscr*, endothelial cell specific chemotaxis receptor was originally identified bioinformatically as an endothelium enriched factor<sup>324</sup>. This receptor has been shown to be required for correct migration of posterior angioblasts to the midline as part of the formation of the ICM<sup>325</sup>. VEGF signalling has been proposed to regulate this

migration. In culture Ecsr proteins were shown to colocalise and potentiate signalling through Kdr<sup>325</sup>.

Nrp1b is a co-receptor for VEGFR and the two share an expression pattern through the developing vasculature of the zebrafish<sup>326</sup>. Upon knockdown of this co-receptor blood vessel patterning was disrupted and crucial arteriovenous connections failed to be formed<sup>326</sup>.

Hspa12b is a member of the heat shock 70 family that is highly restricted to endothelial cells, especially at active sites of vessel sprouting<sup>327</sup>. Hu *et al.* used morpholinos to show that *hspa12b* is required for proper vascular system formation and that this could be due to modulation of the phosphorylation state of the anti-apoptotic protein Akt<sup>327</sup>.

### **Molecular function of 10ss posterior enriched genes over negative controls**

The top over-represented molecular function calculated using the PANTHER database was oxygen transport, which correlated with biological function analysis described above. Histone methyl-transferase activity was also a highly over-represented molecular GO term for the enriched genes of the 10ss posterior *scd*<sup>+</sup> population.

Histone lysine methylation is involved in a range of transcriptional outcomes, and is sensitive to the specific histone type, the lysine position and the number of methyl groups added. Table 4-12 details the significantly enriched genes that are annotated as histone lysine methyltransferases. H3K4 trimethylases of the *kmt2* family feature strongly in this enriched gene list, along with H3K36 methylases and Dot11, the only known H3K79 methylase. The Kmt2 protein family, including the human MLL and drosophila Trithorax proteins, deposit the H3K4me3 mark, which is associated with euchromatin and active transcription<sup>328,329</sup>. Ktm2d knockdown causes widespread

defects including cardiac malformation<sup>330</sup> H3K36me is found in the gene body of actively transcribed genes and is deposited in association with transcriptional elongation<sup>331</sup>. Similarly H3K79me is associated with transcriptional activation<sup>332-334</sup>.

**Table 4-12 Genes annotated as histone lysine-methyl-transferases enriched in the 10ss posterior *scf* expressing population.**

Gene Name	Description	Mean fpkm	log2 FoldChange	p-value	Function
kmt2d	lysine (K)-specific methyltransferase 2D	9.82	1.707	1.54E-05	H3K4 trimethylation
suv420h2	suppressor of variegation 4-20 homolog 2	17.30	1.601	3.69E-05	H3K20 trimethylation
ash1l	ash1 (absent, small, or homeotic)-like	8.82	1.590	1.19E-04	H3K4 trimethylation
dot1l	DOT1-like histone H3K79 methyltransferase	7.19	1.552	1.09E-04	H3 K36 methylation
setd2	SET domain containing 2	18.50	1.518	1.37E-04	H3K4 trimethylation
kmt2cb	lysine (K)-specific methyltransferase 2Cb	9.96	1.345	5.17E-04	H3K4 trimethylation
nsd1a	nuclear receptor binding SET domain protein 1a	20.01	1.310	1.19E-05	H3 K36 methylation
kmt2bb	lysine (K)-specific methyltransferase 2Bb	9.30	1.277	9.32E-04	H3K4 trimethylation
kmt2ca	lysine (K)-specific methyltransferase 2Ca	2.83	1.208	6.99E-03	H3K4 trimethylation
kmt2ba	lysine (K)-specific methyltransferase 2Ba	10.65	1.180	8.82E-04	H3K4 trimethylation
setdb1a	SET domain, bifurcated 1a	34.61	1.124	4.93E-04	H3K4 trimethylation
whsc111	Wolf-Hirschhorn syndrome candidate 1-like 1	13.08	0.841	7.88E-03	H3 K36 methylation
kmt2a	lysine (K)-specific methyltransferase 2A	8.16	0.782	3.98E-02	H3K4 trimethylation
men1	multiple endocrine neoplasia I	32.91	0.750	2.29E-02	Scaffold protein
whsc1	Wolf-Hirschhorn syndrome candidate 1	29.66	0.673	2.62E-02	H3 K36 methylation

Table of histone-lysine-methyl-transferases, identified through biological gene ontology analysis of genes enriched in the 10ss posterior *scf*<sup>+</sup> population upon comparison to expression in the 10ss posterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

### Protein class analysis of 10ss posterior enriched genes over negative controls

8.3% of genes were annotated as transcription factors and 4.2% as signalling molecules, when protein class analysis was carried out on the 10ss *scf*<sup>+</sup> posterior enriched gene list. Biological GO term analysis of the transcription factors,

produced similar top over-represented terms to those of the entire gene list, including vasculogenesis and artery development, however erythroid development is absent while neutrophil differentiation is the top over-represented term. In the zebrafish genome a total of 13 genes are associated with the term neutrophil differentiation, within this enriched gene list four of these genes are featured. These four genes are *etv2*, *spi1b*, *gata1a* and *ncor1*. The first three of these factors have previously been discussed and shown to be key haematopoietic factors contributing to multiple lineages. This suggestion of neutrophil differentiation may be a result of poor data base annotation, as it is not supported by enriched expression of specific neutrophil markers such as *cxcra4*.

Ncor1 is a widely expressed co-repressor identified in anterior-posterior brain patterning through repressing RA signalling<sup>335</sup>. Knock out of *ncor1* revealed that it contributed to neutrophil development, and could be rescued through ectopic *scl/lmo2* expression<sup>336</sup>. In conclusion this highlights one of the key issues with GO analysis, that the annotation are based and limited by the current published literature contained within its databases. Thus a master regulator of a process contributes as much to a GO term as a factor that has been shown to give a mild indirect effect to the process.

### **Top enriched gene lists for the 10ss posterior, compared to negative controls**

39 of the top 50 enriched genes in the *scl*<sup>+</sup> expressing cells of the 10ss posterior, are associated with mapped transcripts in the *scl*<sup>-</sup> sample, thus can be used to calculate meaningful enrichment values. 19 of these top enriched genes have been shown to contribute to haematopoietic or vascular development as detailed in table 4-13. A further 4 of the top enriched genes have been previously identified as enriched in blood or endothelial cell types. The remaining 16 top enriched factors include two

signalling molecules, a chloride channel and 13 genes whose contribution to haematopoietic development has not been assessed (9 lack any functional testing). This is especially exciting as the majority of genes in this list (19 of the 30 studied genes) have been previously demonstrated to play important roles related to haemangiogenesis.

### **10ss posterior depleted genes compared to negative controls**

210 biological GO terms were determined to associate with the 1463 significantly depleted genes in the *sc<sup>+</sup>* cells of the 10ss posterior compared to their non-*sc* expressing neighbours. This included both ubiquitous and tissue-specific processes. The top over-represented term for these depleted factors was mesenchyme morphogenesis, confirming the differentiation of this posterior *sc<sup>+</sup>* population strongly away from certain non-haemangiogenic lineages. Similarly to the 10ss anterior *sc<sup>+</sup>* populations, ribosome biogenesis was dramatically depleted, along with protein translation and ATP synthesis. Negative regulation of neurogenesis was also an over-represented term for the genes depleted in the 10ss posterior *sc<sup>+</sup>* population. This absence of repression for the neural fate was an interesting observation- if this is not an artefact of poor annotation this could be the result of differentiation of the *sc<sup>+</sup>* population to a stage that is no longer competent of switching to a neural fate, thus repression is no longer necessary.

Another possibility is that in other posterior cell types, at the 10ss, are actively repressing the neural fate resulting in the apparent depletion of these factors in *sc<sup>+</sup>* cells. In a similar manner, biological GO terms relating to brain, muscle and somite development were also observed to be depleted in the 10ss posterior *sc<sup>+</sup>* population. Genes annotated with immune and cardiac development GO terms were also identified as significantly depleted in the 10ss posterior *sc<sup>+</sup>* population, despite several of the key cardiac and myeloid terms being over-represented in the enriched

genes of the 10ss anterior *scf*<sup>+</sup> population (Figure 4-16). This further suggests that at the 10ss the anterior and posterior *scf*<sup>+</sup> populations have already acquired different developmental fates.

**Table 4-13 Top enriched genes in the 10ss posterior *scf* expressing population.**

Gene Name	Description	Mean fpm	log2 FoldChange	p-value	Novel?
grpa	GRB2-related adaptor protein a	32.86	9.582	1.42E-28	
agtr2	angiotensin II receptor, type 2	21.27	8.977	1.06E-22	
ecscr	endothelial cell surface expressed chemotaxis and apoptosis regulator	6.48	8.553	1.00E-20	
morc3b	MORC family CW-type zinc finger 3b	471.48	7.985	2.24E-138	
VAV3	Vav3	7.96	7.621	3.40E-17	
gata1a	GATA binding protein 1a	634.37	7.610	1.84E-116	
si:ch73-248e21.5	Novel	4.22	7.370	7.61E-14	NOVEL
fgd5a	FYVE, RhoGEF and PH domain containing 5a	14.91	7.256	3.04E-40	
gf11aa	growth factor independent 1A transcription repressor a	1861.89	7.253	5.30E-70	
adcyap1r1b	adenylate cyclase activating polypeptide 1b (pituitary) receptor type I	16.49	7.184	6.04E-37	
rps6ka3b	ribosomal protein S6 kinase, polypeptide 3b	24.27	7.129	1.63E-72	NOVEL
dnase1l3l	deoxyribonuclease I-like 3, like	7.91	7.084	1.57E-12	
clie2	chloride intracellular channel 2	45.05	7.071	6.63E-31	
kdr	kinase insert domain receptor (a type III receptor tyrosine kinase)	16.80	7.017	3.95E-41	
HBZ	hemoglobin zeta	12.24	7.011	3.43E-12	
arhgap27	Rho GTPase activating protein 27	2.05	6.997	3.59E-12	
cx45.6	connexin 45.6	12.49	6.986	3.06E-26	
tie1	tyrosine kinase with immunoglobulin-like and EGF-like domains 1	21.86	6.926	2.90E-32	
fl1b	Fli-1 proto-oncogene, ETS transcription factor b	43.94	6.896	4.47E-68	
flt4	fms-related tyrosine kinase 4	36.26	6.868	1.44E-75	
PAQR9	progesterin and adipoQ receptor family member IX	6.94	6.749	3.87E-11	NOVEL
exoc3l2b	exocyst complex component 3-like 2b	2.37	6.717	3.49E-17	
stab2	stabilin 2	23.52	6.699	1.36E-30	
ral1	T-cell acute lymphocytic leukemia 1	627.24	6.699	3.87E-45	
ACER2	alkaline ceramidase 2	29.75	6.697	1.60E-21	NOVEL
si:dkey-183p4.9	Novel	9.66	6.675	7.05E-11	NOVEL
si:ch73-315i10.1	Novel	3.73	6.572	1.80E-10	NOVEL
esama	endothelial cell adhesion molecule a	41.08	6.529	1.35E-45	
htf1fb	5-hydroxytryptamine (serotonin) receptor 1Fb	8.35	6.428	1.70E-11	NOVEL
lmo2	LIM domain only 2 (rhotbotin-like 1)	3763.49	6.369	1.22E-52	
fl1a	Fli-1 proto-oncogene, ETS transcription factor a	304.34	6.321	1.02E-66	
klf17	Kruppel-like factor 17	1134.43	6.297	3.40E-107	
si:ch211-14k19.8	Novel	8.70	6.204	1.06E-22	NOVEL
she	Src homology 2 domain containing E	37.46	6.203	1.58E-40	
scarf1	scavenger receptor class F, member 1	22.52	6.195	1.26E-46	
sox7	SRY (sex determining region Y)-box 7	386.91	6.189	1.36E-64	
exoc3l1	exocyst complex component 3-like 1	32.35	6.161	2.81E-40	
sptb	spectrin, beta, erythrocytic	17.79	6.095	4.16E-08	
isl1l	islet1, like	3.01	5.977	9.54E-10	
aqp8a.1	aquaporin 8a, tandem duplicate 1	29.64	5.923	9.88E-30	
AL929291.1	Novel	3.83	5.917	2.94E-08	NOVEL
si:dkey-261j4.3	Novel	40.54	5.886	1.63E-34	NOVEL
nrp1b	neuropilin 1b	38.29	5.870	9.66E-61	
tns2a	tensin 2a	13.45	5.866	2.02E-34	
btk	Bruton agammaglobulinemia tyrosine kinase	62.90	5.842	1.02E-58	
rell2	RELT-like 2	7.12	5.823	2.03E-13	NOVEL
tcf21	transcription factor 21	11.84	5.807	6.54E-13	
tfri1a	transferrin receptor 1a	27.97	5.792	1.14E-19	
drl	draculin	74.71	5.787	1.28E-50	
si:dkey-207j16.5	Novel	4.68	5.754	1.86E-08	NOVEL

Table of the top 50 enriched factors of the 10ss posterior *scf*<sup>+</sup> population upon comparison to expression in the 10ss posterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

## Conclusions

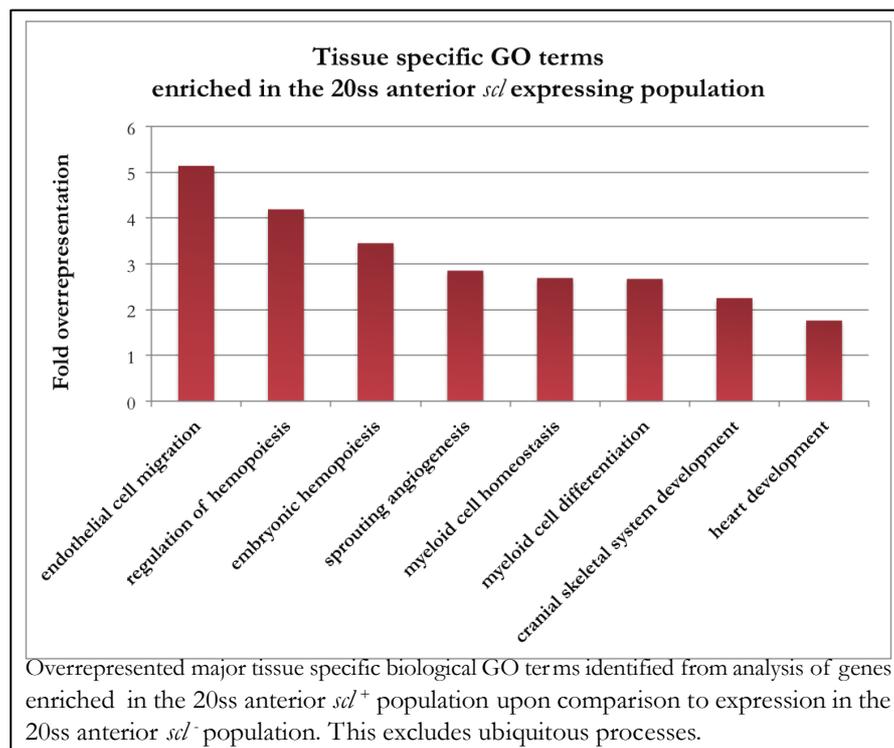
The 10ss posterior *scf* expressing population shows specific enrichment of genes correlating strongly with erythroid development and function in addition to blood vessel morphogenesis. This correlates with previously suggested roles for the

posterior lateral mesoderm angioblasts. Ras signalling has been highly studied in haematopoiesis using transformed culture models with a range of conclusions that are highly cell line specific. Biological GO analysis of this transcriptomic data also highlighted Ras signalling as an over-represented process indicating that the Ras pathway could potentially mediate these developmental processes.

The enriched factors for erythroid and blood vessel remodelling have been identified in previous studies confirming the validity of the RNA-seq technique to probe these early developmental populations. However GO enrichment analysis struggles with annotation especially for vascular functions and can give different over-representation results based on genes that are associated with a range of vascular functions. As with the 10ss anterior *scf*<sup>+</sup> population, the significantly enriched gene list includes not only previously studied proteins but novel factors too.

#### 4.5.4. 20ss Anterior

**Figure 4-19 Over-represented tissue-specific gene ontology terms for genes enriched in the 20ss anterior *scf* expressing population.**



10601 genes showing significantly different expression were identified through DESeq2 analysis of the RNA-seq results for the 20ss anterior *scf*<sup>+</sup> versus *scf*<sup>-</sup> populations. This significant increase in the number of differentially expressed genes could mean that these populations become more divergent and differentiated with development to the 20ss.

### **20ss anterior enriched genes over negative controls**

4274 genes were significantly enriched in the 20ss anterior *scf*<sup>+</sup> cells compared to surrounding *scf*<sup>-</sup> expressing cells. 141 over-represented biological GO terms were associated with these enriched genes and 896 unclassified genes were identified. 18 tissue-specific processes were among the over-represented biological GO terms— as with the 10ss enriched analyses, all were haematopoietic or vascular functions, many of which were maintained from the 10ss populations (See figure 4-19).

#### Gas transport

14 genes contributed to the over-represented GO term gas transport- 12 globin genes plus carbonic anhydrase and aquaporin1a were significantly enriched in the *scf*<sup>+</sup> cells of the 20ss anterior (table 4-14). The enriched globin genes included subunits for embryonic haemoglobin alpha, beta & zeta proteins plus their different isoforms and subunits.

Aquaporin1a has two homologues in zebrafish, both of which are enriched the 20ss anterior *scf*<sup>+</sup> cells, and expressed through the developing vasculature and erythrocytes<sup>337,338</sup>. This transporter is capable of both water and gas transport, showing CO<sub>2</sub> and NH<sub>3</sub> permeability<sup>337</sup> and is downstream of *etv2* and *gata1* signalling<sup>338</sup>. Overexpression of *aqp1* in erythroleukemic cell culture was shown to promote erythroid differentiation, including haemoglobin expression<sup>339,340</sup>. Carbonic anhydrase expression has been shown to increase during zebrafish development<sup>341</sup>

with expression in haematopoietic and vascular cells and lie downstream of the *cloche* mutation<sup>342</sup>.

**Table 4-14 Genes annotated with the GO term oxygen transport enriched in the 20ss anterior *scf* expressing population.**

Gene name	Description	Mean fpkm	log2 FoldChange	p-value
aqp1a.2	aquaporin 1a (Colton blood group), tandem duplicate 2	15.30	9.396	9.14E-17
hbae1	hemoglobin, alpha embryonic 1	33.52	8.739	1.12E-20
HBZ	zgc:163057	23.28	7.335	7.71E-23
hbbe2	hemoglobin beta embryonic-2	75.68	6.461	3.93E-42
hbbe1.1	hemoglobin beta embryonic-1.1	199.46	6.083	3.10E-43
hbae0	hemoglobin, alpha embryonic 1	286.75	6.069	5.65E-15
hbz	hemoglobin zeta	7.38	5.808	7.98E-18
hbbe3	hemoglobin beta embryonic-3	658.27	5.721	2.26E-50
hbae3	hemoglobin alpha embryonic-3	676.62	5.468	4.35E-14
aqp1a.1	aquaporin 1a (Colton blood group), tandem duplicate 1	243.70	5.058	1.57E-53
hbbe1.2	hemoglobin beta embryonic-1.2	13.92	5.016	3.70E-16
hbbe1.1	hemoglobin beta embryonic-1.1	202.31	4.215	6.19E-29
hbae1	hemoglobin, alpha embryonic 1	4.66	4.037	8.26E-09
cahz	carbonic anhydrase	45.08	3.453	3.70E-04

Table of oxygen transporter molecules identified by biological GO analysis of genes enriched in the 20ss anterior *scf*<sup>+</sup> population upon comparison to expression in the 20ss anterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

The enriched 10ss posterior *scf*<sup>+</sup> population also identified oxygen transport as an over-represented biological GO term. The expression levels of the haemoglobin genes *hbbe1*, *hbbe2*, *hbbe3*, *hbae1* and *hbae3*, are significantly greater in the 20ss anterior than in the 10ss posterior, despite erythroid development mainly occurring in the posterior of zebrafish embryos. This finding indicates that the embryonic location in which erythroid development occurs, has expanded anteriorly and that this expression is strong by the 20ss. It would be interesting to identify the location or trace the movement of these anterior erythroid cells to confirm the spreading of the posterior erythroid population or identify *de novo* erythroid sites of anterior development.

#### Erythrocyte differentiation

The biological GO term erythroid differentiation was also over-represented in the 20ss posterior *scf* expressing cells. This GO term was associated with 23 enriched

genes, which are detailed in Table 4-15, these included *gata1a*, *gata2a*, *lmo2* and *alas2*.

Other enriched genes included *slc4a1a* and *epb41b*.

Slc4a1a is a solute carrier protein that upon mutation results in the sudden loss of circulating cells after the first 3 days of zebrafish development<sup>276</sup>. This solute carrier is also an erythrocyte specific cytoskeletal protein that is required for mitosis of erythroid cells<sup>343</sup>. The loss of erythroid cells in *slc4a1a* mutants was shown to be a result of apoptosis of erythroid cells that had failed to complete mitosis and cytokinesis<sup>343</sup>. Epb41b mutation produced a similar phenotype as observed for *slc4a1a* knockdown- phenotypically normal onset of circulation followed by sudden severe anaemia at ~3dpf<sup>276,344</sup>. The Epb41b protein is required to anchor the cytoskeleton of erythroid cells to the plasma membrane, upon mutation membrane defects are observed along with increased osmotic fragility<sup>345</sup>.

**Table 4-15 Genes annotated with the GO term erythroid differentiation enriched in the 20ss anterior *slc* expressing population.**

Gene Name	Description	Mean FPKM	log2 Fold Change	p-value
tal1	T-cell acute lymphocytic leukemia 1	883.64	9.084	0.0E+00
spi1b	Spi-1 proto-oncogene b	201.04	8.781	2.2E-302
lmo2	LIM domain only 2 (rhombotin-like 1)	1312.93	8.611	0.0E+00
gata1a	GATA binding protein 1a	50.12	8.167	6.3E-13
slc4a1a	solute carrier family 4, member 1a	21.01	6.775	1.6E-67
sptb	spectrin, beta, erythrocytic	26.77	6.502	4.4E-16
alas2	aminolevulinate, delta-, synthase 2	38.25	6.308	1.5E-14
gf11a	growth factor independent 1A transcription repressor a	90.34	5.298	2.4E-08
epb41b	erythrocyte membrane protein band 4.1b	8.34	5.078	1.7E-06
slc25a37	solute carrier family 25, member 37	18.83	3.355	7.9E-04
rasa3	RAS p21 protein activator 3	48.82	3.159	1.5E-31
numb	numb homolog (Drosophila)	9.46	2.933	6.1E-30
gata2a	GATA binding protein 2a	31.02	2.318	5.4E-15
tmod2	tropomodulin 2	38.28	2.218	2.3E-31
fech	ferrochelatase	15.89	2.195	1.5E-08
slc25a38a	solute carrier family 25, member 38a	3.79	1.865	1.3E-02
stat5a	signal transducer and activator of transcription 5a	6.30	1.463	5.0E-07
vhl	von Hippel-Lindau tumor suppressor	11.86	1.400	1.9E-06
sec23b	Sec23 homolog B, COPII coat complex component	25.04	1.135	9.1E-12
jak2b	Janus kinase 2b	7.01	0.607	5.0E-03
tert	telomerase reverse transcriptase	11.82	0.560	2.5E-02
adnp2a	ADNP homeobox 2a	8.46	0.530	4.2E-02
adnp2b	ADNP homeobox 2b	14.96	0.473	7.6E-03
atp1f1b	ATPase inhibitory factor 1b	77.03	0.356	4.6E-02

Table of factors involved in erythroid differentiation, identified through biological gene ontology analysis of genes enriched in the 20ss anterior *slc*<sup>+</sup> population upon comparison to expression in the 20ss anterior *slc*<sup>-</sup> population. Factors are in order of fold enrichment.

The enrichment of these genes suggests that by the 20ss primitive erythroid cells have developed, initiated high levels of haem synthesis and begun to proliferate in preparation for the initiation of circulation, even in the anterior *scf*<sup>+</sup> population, which only minimally contributes to the erythroid pool.

### Embryonic haematopoiesis

The GO term for embryonic haematopoiesis is over-represented in the enriched gene list for the 20ss anterior *scf*<sup>+</sup> population and contains key haematopoietic factors as listed in table 4-16.

**Table 4-16 Genes annotated with the GO term embryonic haematopoiesis enriched in the 20ss anterior *scf* expressing population.**

Gene Name	Description	Mean FPKM	log2 Fold Change	p-value
<i>etv2</i>	ets variant 2	<b>1384.75</b>	9.289	0.0E+00
<i>tal1</i>	T-cell acute lymphocytic leukemia 1	<b>883.64</b>	9.084	0.0E+00
<i>lmo2</i>	LIM domain only 2 (rhombotin-like 1)	<b>1312.93</b>	8.611	0.0E+00
<i>erg</i>	v-ets avian erythroblastosis virus E26 oncogene homolog	<b>86.16</b>	8.257	2.2E-158
<i>gata1a</i>	GATA binding protein 1a	<b>50.12</b>	8.167	6.3E-13
<i>myct1a</i>	myc target 1a	<b>10.78</b>	7.930	8.2E-25
<i>csf3r</i>	colony stimulating factor 3 receptor (granulocyte)	<b>8.61</b>	7.117	7.3E-34
<i>slc4a1a</i>	solute carrier family 4, member 1a	<b>21.01</b>	6.775	1.6E-67
<i>sptb</i>	spectrin, beta, erythrocytic	<b>26.77</b>	6.502	4.4E-16
<i>alas2</i>	aminolevulinate, delta-, synthase 2	<b>38.25</b>	6.308	1.5E-14
<i>cebpa</i>	CCAAT/enhancer binding protein (C/EBP), alpha	<b>208.21</b>	5.702	1.5E-126
<i>tll1</i>	tolloid-like 1	<b>5.12</b>	5.675	2.3E-41
<i>tmem88a</i>	transmembrane protein 88 a	<b>331.78</b>	5.671	1.8E-115
<i>gfi1aa</i>	growth factor independent 1A transcription repressor a	<b>90.34</b>	5.298	2.4E-08
<i>epb41b</i>	erythrocyte membrane protein band 4.1b	<b>8.34</b>	5.078	1.7E-06
<i>irak3</i>	interleukin-1 receptor-associated kinase 3	<b>4.70</b>	5.031	4.2E-24
<i>snrkb</i>	SNF related kinase b	<b>15.12</b>	4.198	4.2E-38
<i>csrnp1a</i>	cysteine-serine-rich nuclear protein 1a	<b>15.23</b>	4.139	8.3E-15
<i>slc25a37</i>	solute carrier family 25, member 37	<b>18.83</b>	3.355	7.9E-04
<i>numb</i>	numb homolog (Drosophila)	<b>9.46</b>	2.933	6.1E-30
<i>tbx20</i>	T-box 20	<b>5.90</b>	2.375	6.2E-12
<i>gata2a</i>	GATA binding protein 2a	<b>31.02</b>	2.318	5.4E-15
<i>socs1a</i>	suppressor of cytokine signaling 1a	<b>4.07</b>	1.583	1.4E-03
<i>dhx8</i>	DEAH (Asp-Glu-Ala-His) box polypeptide 8	<b>20.40</b>	0.663	2.7E-03
<i>prkcbb</i>	protein kinase C, beta b	<b>4.81</b>	0.581	2.5E-02
<i>mta3</i>	metastasis associated 1 family, member 3	<b>44.61</b>	0.531	5.3E-03
<i>acvr11</i>	activin A receptor, type I like	<b>20.40</b>	0.384	1.8E-02

Table of factors involved in embryonic haematopoiesis, identified through biological gene ontology analysis of genes enriched in the 20ss anterior *scf*<sup>+</sup> population upon comparison to expression in the 20ss anterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

These include general haematopoietic transcription factors *etv2* and *lmo2*, plus key erythroid and vascular factors, which have previously been discussed. In addition to

these *myct1a*, *csf3r* and *c/ebpα* genes are also specifically enriched in the *scf*<sup>+</sup> population of the 20ss anterior. *Myct1a* is a putative *myc* target gene that has been shown to play a role in HSC fate decisions due to morpholino knockdown causing both myeloid and erythroid defects<sup>346</sup>. *Csf3r* is a granulocyte colony stimulating factor that is required for the development of anterior myeloid cells during primitive haematopoiesis, plus initiates emergency haematopoiesis in response to bacterial infection<sup>280</sup>. Knockdown of this factor dramatically reduced *runx1* expression<sup>347</sup> and signalling through the granulocyte colony stimulating pathway is required for HSPC development and proliferation<sup>348</sup>.

*C/ebpα* is a transcription factor that is highly conserved among vertebrates and binds the DNA sequence CCAAT<sup>349</sup>. Studied in other models has shown that *C/ebpα* acts as a key granulocyte differentiation factor<sup>349-351</sup>. In zebrafish *c/ebpα* is expressed in a pattern matching *scf*, suggesting that these transcription factors may inter-regulate<sup>352</sup>. In other model systems *c/ebpα* was shown to favour the myeloid fate by repressing primitive erythropoiesis<sup>352</sup>.

The biological GO term embryonic haematopoiesis was also over-represented in the 10ss anterior. As this anterior *scf*<sup>+</sup> population has progressed from the 10 to 20ss- a greater number of embryonic haematopoiesis associated factors are enriched potentially as a result in an increase in complexity of this *scf*<sup>+</sup> population. Comparison of these enriched genes shows the maintained expression and enrichment in the *scf*<sup>+</sup> population of vascular factors (*tbx20* and *ery*), general haematopoietic regulators (*etv2* and *lmo2*) plus key erythroid genes (*gata1*, *gfi1aa*, *sptb* and *alas1*). The continued enriched expression of these factors within the anterior *scf*<sup>+</sup> population suggests that this population maintains features of three key haemangiogenic lineages. By the 20ss these lineages are divided between distinct *scf*<sup>+</sup> sub-populations that I will describe in the next chapter. This raises the question of

whether these lineages are already distinct at the 10ss or that these factors are co-expressed within individual cells that later diverge into separate subpopulations.

#### Myeloid cell differentiation

The myeloid lineage has been shown to initially arise from the anterior lateral plate mesoderm, before granulocytes and macrophages disperse through the embryo, eventually residing in the kidney and circulating blood. The analysis of the *scf*<sup>+</sup> population at the 10ss indicated a myeloid profile in agreement with these previous studies. Biological GO term over-representation analysis of the 20ss anterior *scf*<sup>+</sup> enriched gene list determined the myeloid cell differentiation term as over-represented by 2.83 fold. 42 genes contributed directly to this process from the enriched gene list, these are recorded in table 4-17. *etv2*, *lmo2*, *gata5* and *spi1b* showed particularly high enrichment over expression levels in neighbouring tissues however their expression level was significantly reduced in comparison to the 10ss anterior *scf*<sup>+</sup> population. This suggested that these key myeloid factors were highly specifically expressed within this anterior *scf*<sup>+</sup> population during development, yet at lower levels or in a smaller percent of this *scf*<sup>+</sup> population at the 20ss. Biologically this could mean that the 20ss *scf*<sup>+</sup> population is diverting from the myeloid lineage or that the Citrine<sup>+</sup> cells of the 20ss anterior represent a collection of *scf*<sup>+</sup> sub-populations, of which at least one expressed myeloid factors.

Consistent with the latter, *c/ebpα* (key granulocyte differentiation factor) expression was seen to double between the 10ss and 20ss anterior *scf*<sup>+</sup> samples, suggesting the continued and progressive development of the myeloid lineage within the *scf*<sup>+</sup> population.

**Table 4-17 Genes annotated with the GO term myeloid differentiation enriched in the 20ss anterior *scl* expressing population.**

Gene Name	Description	Mean FPKM	log2 Fold Change	p-value
etv2	ets variant 2	1384.75	9.289	0.0E+00
tal1	T-cell acute lymphocytic leukemia 1	883.64	9.084	0.0E+00
spi1b	Spi-1 proto-oncogene b	201.04	8.781	2.2E-302
lmo2	LIM domain only 2 (rhombotin-like 1)	1312.93	8.611	0.0E+00
gata1a	GATA binding protein 1a	50.12	8.167	6.3E-13
csf3r	colony stimulating factor 3 receptor (granulocyte)	8.61	7.117	7.3E-34
slc4a1a	solute carrier family 4, member 1a	21.01	6.775	1.6E-67
sptb	spectrin, beta, erythrocytic	26.77	6.502	4.4E-16
alas2	aminolevulinate, delta-, synthase 2	38.25	6.308	1.5E-14
irf8	interferon regulatory factor 8	20.21	6.295	1.4E-72
tmem88a	transmembrane protein 88 a	331.78	5.671	1.8E-115
gfi1aa	growth factor independent 1A transcription repressor a	90.34	5.298	2.4E-08
epb41b	erythrocyte membrane protein band 4.1b	8.34	5.078	1.7E-06
slc25a37	solute carrier family 25, member 37	18.83	3.355	7.9E-04
gata5	GATA binding protein 5	28.95	3.248	2.4E-43
casp8	caspase 8, apoptosis-related cysteine peptidase	6.95	3.236	3.9E-25
rasa3	RAS p21 protein activator 3	48.82	3.159	1.5E-31
gfi1ab	growth factor independent 1A transcription repressor b	29.19	3.124	9.8E-30
numb	numb homolog (Drosophila)	9.46	2.933	6.1E-30
gata2a	GATA binding protein 2a	31.02	2.318	5.4E-15
tmod2	tropomodulin 2	38.28	2.218	2.3E-31
fech	ferrochelatase	15.89	2.195	1.5E-08
kif1b	kinesin family member 1B	9.53	2.061	2.9E-31
slc25a38a	solute carrier family 25, member 38a	3.79	1.865	1.3E-02
plcg1	phospholipase C, gamma 1	45.58	1.789	1.7E-22
stat5a	signal transducer and activator of transcription 5a	6.30	1.463	5.0E-07
vhl	von Hippel-Lindau tumor suppressor	11.86	1.400	1.9E-06
akap10	A kinase (PRKA) anchor protein 10	8.86	1.305	8.3E-06
smad9	SMAD family member 9	7.60	1.234	1.9E-07
sec23b	Sec23 homolog B, COPII coat complex component	25.04	1.135	9.1E-12
ptenb	phosphatase and tensin homolog B	61.99	1.128	7.2E-14
npc1	Niemann-Pick disease, type C1	18.25	0.970	4.2E-09
ncor2	nuclear receptor corepressor 2	22.29	0.925	1.0E-08
jagn1b	jagunal homolog 1b	26.37	0.649	8.6E-04
jak2b	Janus kinase 2b	7.01	0.607	5.0E-03
brf1b	BRF1, RNA polymerase III transcription initiation factor b	9.35	0.599	2.6E-02
ncor1	nuclear receptor corepressor 1	17.33	0.562	6.4E-04
tert	telomerase reverse transcriptase	11.82	0.560	2.5E-02
adnp2a	ADNP homeobox 2a	8.46	0.530	4.2E-02
wasla	Wiskott-Aldrich syndrome-like a	5.50	0.492	4.9E-02
adnp2b	ADNP homeobox 2b	14.96	0.473	7.6E-03
brd2a	bromodomain containing 2a	43.45	0.390	8.3E-03
atpif1b	ATPase inhibitory factor 1b	77.03	0.356	4.6E-02

Table of myeloid factors, expressed and enriched in the 20ss anterior *scl*<sup>+</sup> population upon comparison to expression in the 20ss anterior *scl*<sup>-</sup> population. Factors are in order of fold enrichment.

### Endothelial cell migration

12 enriched genes correlated with the biological GO term endothelial cell migration, which only has 20 zebrafish genes assigned to it. This strong over-representation is a

result of small gene list size, and the fact that it is populated by factors annotated with multiple vascular functions. However the suggestion that *scf*<sup>+</sup> angioblasts of the anterior undergo significant movement as the key vessels of the cranial vasculature are formed is supported by time-lapse imaging (Figure 3-9). These enriched genes are listed in Table 4-18 and include *plekhg5a*, *fn1a*, *dll4* and *itga5*.

**Table 4-18 Genes annotated with the GO term endothelial cell migration enriched in the 20ss anterior *scf* expressing population.**

Gene Name	Description	Mean fpkm	log2 FoldChange	p-value
<i>plekhg5a</i>	pleckstrin homology domain containing, family G (with RhoGef domain) member 5a	33.03	4.896	3.69E-165
<i>fn1a</i>	fibronectin 1a	92.18	4.031	1.19E-107
<i>dll4</i>	delta-like 4 (Drosophila)	84.20	4.030	4.77E-81
<i>itga5</i>	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	83.72	3.977	1.29E-38
<i>cxcr4a</i>	chemokine (C-X-C motif) receptor 4a	44.05	2.987	2.52E-27
<i>vegfc</i>	vascular endothelial growth factor c	14.61	2.422	3.38E-26
<i>rab13</i>	RAB13, member RAS oncogene family	41.48	1.572	1.67E-08
<i>amot</i>	angiominin	7.58	1.394	3.88E-12
<i>hspg2</i>	heparan sulfate proteoglycan 2	13.76	1.374	1.20E-18
<i>efnb2a</i>	ephrin-B2a	40.39	1.293	1.12E-17
<i>robo4</i>	roundabout, axon guidance receptor, homolog 4 (Drosophila)	35.60	0.669	2.37E-05
<i>cxcl12b</i>	chemokine (C-X-C motif) ligand 12b (stromal cell-derived factor 1)	51.38	0.509	4.68E-03

Table of factors involved in endothelial cell migration, identified through biological gene ontology analysis of genes enriched in the 20ss anterior *scf*<sup>+</sup> population upon comparison to expression in the 20ss anterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

*fn1a* was identified through a mutation that caused two hearts to form in zebrafish<sup>353</sup>. The *fn1a* gene encodes fibronectin and was shown to be required for correct migration of myocardial precursors to the midline of the embryo<sup>354</sup>. Chiu *et al.* demonstrated that *fn1a* was required for endothelial migration and tissue invasion processes, and that this process required *itga5* (Integrin- $\alpha$ 5)<sup>355</sup>, as knockdown of either *fn1a* or *itga5* produced the same defects in angiogenic sprouting. In vivo imaging and specific knockdown models would be required to conclude which of these developmental processes *fn1a* was contributing to in the 20ss anterior *scf*<sup>+</sup> population.

Delta-like 4 is a Notch signalling ligand that is required for regulating endothelial cell migration and proliferation, thus counteracting VEGF signalling, to limit

vascularisation<sup>356</sup>. Complete loss of *dll4* results in major arterial defects<sup>357</sup> despite artery differentiation being unaffected in *dll4* deficiency<sup>358</sup>.

### Angiogenesis

Concurrent with continued and significant development of the circulatory system, angiogenic terms were also over-represented biological GO process. Over double the number of genes contributed to this term in the 20ss anterior *scf*<sup>+</sup> population than at the 10ss, as detailed in table 4-19. Enriched expression was maintained of all 10ss anterior angiogenic factors, all of which showed an increase in expression with the exception of *etv2* and *fn1a*. Top enriched angiogenic factors in the 20ss *scf*<sup>+</sup> anterior population show over 100 fold enrichment compared to neighbouring citrine<sup>-</sup> cells, which suggests high specificity and restriction of the vascular fate to the *scf*<sup>+</sup> population. Together this data described a population that is highly active in vascular development and has considerably increased capacity for blood vessel formation compared to the same population at the 10ss. This enriched gene list also included genes that contribute to cardiac development through their role in regulating migration of myocytes or endocardial cells into the developing heart field<sup>294,307,355</sup>. The enriched expression of factors that contribute to the formation of the heart, in the absence of cardiac markers such as *hand2* and *myl7*<sup>141,359</sup>, could suggest that this 20ss anterior *scf*<sup>+</sup> population is indirectly involved in cardiac development. This would suggest that within the endothelial compartment of the 20ss anterior *scf*<sup>+</sup> population, distinct subpopulations may exist which participate in either cranial vasculature or the developing heart.

**Table 4-19 Genes annotated with the GO term angiogenic factors enriched in the 20ss anterior *scf* expressing population.**

Gene Name	Description	Mean FPKM	log2 Fold Change	p-value
etv2	ets variant 2	1384.75	9.289	0.0E+00
cdh5	cadherin 5	116.26	9.134	0.0E+00
kdr	kinase insert domain receptor	38.59	9.131	1.2E-157
tal1	T-cell acute lymphocytic leukemia 1	883.64	9.084	0.0E+00
erg	v-ets avian erythroblastosis virus E26 oncogene homolog	86.16	8.257	2.2E-158
arhgef9b	Cdc42 guanine nucleotide exchange factor (GEF) 9b	33.38	7.684	3.2E-150
flt1	fms-related tyrosine kinase 1	19.24	7.668	1.8E-144
si:ch73-334d15.2	si:ch73-334d15.2	25.04	7.083	7.0E-53
fli1a	Fli-1 proto-oncogene, ETS transcription factor a	459.33	5.846	4.1E-276
hlx1	H2.0-like homeo box 1 (Drosophila)	62.44	5.586	1.4E-137
mcamb	melanoma cell adhesion molecule b	55.30	5.561	7.0E-187
nrp1b	neuropilin 1b	64.28	5.367	4.9E-194
dab2	Dab, mitogen-responsive phosphoprotein, homolog 2	18.88	4.965	4.1E-82
plekhg5a	pleckstrin homology domain containing, family G member 5a	33.03	4.896	3.7E-165
hspa12b	heat shock protein 12B	26.78	4.894	1.4E-113
ramp2	receptor (G protein-coupled) activity modifying protein 2	111.04	4.399	6.0E-63
fn1a	fibronectin 1a	92.18	4.031	1.2E-107
dll4	delta-like 4 (Drosophila)	84.20	4.030	4.8E-81
itga5	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	83.72	3.977	1.3E-38
hapln1b	hyaluronan and proteoglycan link protein 1b	17.65	3.822	2.4E-54
lgals2a	lectin, galactoside-binding, soluble, 2a	15.06	3.780	3.5E-12
arhgef7b	Rho guanine nucleotide exchange factor (GEF) 7b	98.29	3.643	6.9E-105
fmnl3	formin-like 3	41.20	3.592	3.7E-105
cxcr4a	chemokine (C-X-C motif) receptor 4a	44.05	2.987	2.5E-27
vegfc	vascular endothelial growth factor c	14.61	2.422	3.4E-26
lpar2a	lysophosphatidic acid receptor 2a	37.76	2.354	1.3E-24
cdc42	cell division cycle 42	194.64	1.902	1.0E-32
nrp2b	neuropilin 2b	51.56	1.879	3.0E-22
fermt2	fermitin family member 2	59.01	1.844	2.7E-30
git1	G protein-coupled receptor kinase interacting ArfGAP 1	18.08	1.818	8.8E-25
plcg1	phospholipase C, gamma 1	45.58	1.789	1.7E-22
cds2	CDP-diacylglycerol synthase 2	50.03	1.744	2.9E-19
rab5c	RAB5C, member RAS oncogene family	70.64	1.707	1.6E-13
shc1	SHC transforming protein 1	22.45	1.673	2.6E-21
pld1a	phospholipase D1a	6.34	1.652	1.8E-11
pik3c2a	phosphatidylinositol-4-phosphate 3-kinase, subunit type 2 alpha	8.16	1.594	1.4E-12
rab13	RAB13, member RAS oncogene family	41.48	1.572	1.7E-08
gng2	guanine nucleotide binding protein (G protein), gamma 2	85.62	1.560	5.9E-21
bmpr2a	bone morphogenetic protein receptor, type II a	5.11	1.553	7.1E-10
amotl2a	angiomin like 2a	52.67	1.516	1.3E-18
sat1b	spermidine/spermine N1-acetyltransferase 1b	14.66	1.427	1.6E-05
amot	angiomin	7.58	1.394	3.9E-12
hspg2	heparan sulfate proteoglycan 2	13.76	1.374	1.2E-18
efnb2a	ephrin-B2a	40.39	1.293	1.1E-17
gna13b	guanine nucleotide binding protein (G protein), alpha 13b	21.73	1.251	1.4E-05
sdc2	syndecan 2	27.56	1.196	2.2E-09
hdac6	histone deacetylase 6	7.92	1.179	5.6E-07
rab11a	RAB11a, member RAS oncogene family	51.57	1.143	5.6E-13
hexb	hexosaminidase B (beta polypeptide)	35.74	0.881	7.2E-05
tnnt2a	troponin T type 2a (cardiac)	21.52	0.782	1.4E-03
lpar6a	lysophosphatidic acid receptor 6a	9.32	0.757	8.9E-03
robo4	roundabout, axon guidance receptor, homolog 4 (Drosophila)	35.60	0.669	2.4E-05
ppp1cab	protein phosphatase 1, catalytic subunit, alpha isozyme b	117.47	0.514	4.8E-04
cxcl12b	chemokine (C-X-C motif) ligand 12b	51.38	0.509	4.7E-03
wnk1a	WNK lysine deficient protein kinase 1a	10.44	0.497	3.6E-02
rbpj	recombination signal binding protein for Ig kappa J region a	12.03	0.455	2.1E-02

Table of factors involved in angiogenesis, identified through biological gene ontology analysis of genes enriched in the 20ss anterior *scf*<sup>+</sup> population upon comparison to expression in the 20ss anterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

### **Molecular function of 20ss anterior enriched genes over negative controls**

Oxygen transport was the top over-represented molecular function calculated using the PANTHER database annotations for the 20ss anterior *scf*<sup>+</sup> population, which correlated with the gain of erythroid features to the profile of this anterior population. Rho signalling was also a highly over-represented molecular GO term, and included factors showing guanyl-nucleotide exchange, GTPase activator and GTP binding functions. This could denote the continued use of Rho signalling in the regulation of anterior haemangiogenic fates from the 10ss to the 20ss.

### **Protein class analysis of 20ss anterior enriched genes over negative controls**

7.4% of enriched genes in the 20ss anterior *scf*<sup>+</sup> samples were annotated by the PANTHER database as transcription factors, with 5.5% of genes labelled as signalling molecules. Biological GO term analysis of the transcription factors, showed high over-representation for myeloid differentiation and its regulation, blood vessel development and morphogenesis plus erythrocyte development.

### **Top enriched genes in the 20ss anterior, compared to negative controls**

46 of the top 50 enriched genes had mapped transcripts in the *scf*<sup>+</sup> sample, for the 20ss anterior allowing enrichment values to be calculated. 24 of these top enriched genes, listed in table 4-20, have published experimental data describing their haematopoietic or vascular roles. A further 11 highly enriched genes are known to be expressed within the developing circulatory system of zebrafish or have haemangiogenic roles reported from other models.

The entire remaining top enriched factors (11) for the 20ss *scf*<sup>+</sup> anterior population are novel i.e. any biological or molecular function is predicted from structure, plus

there is no published literature available on the gene. These functionally unstudied and novel genes would be of interest to investigate further, for their potential roles in development of the circulatory system.

**Table 4-20 Top enriched genes in the 20ss anterior *scf* expressing population.**

Gene Name	Description	Mean FPKM	log2 FoldChange	p-value	Novel?
lmo2	LIM domain only 2 (rhombotin-like 1)	1312.93	8.611	0	
sox7	SRY (sex determining region Y)-box 7	730.47	8.026	0	
ets2	ets variant 2	1384.75	9.289	0	
tal1	T-cell acute lymphocytic leukemia 1	883.64	9.084	0	
clec14a	C-type lectin domain family 14, member A	393.02	9.216	4.82E-239	
kdlr	kinase insert domain receptor like	185.31	8.288	2.98E-234	
spi1b	Spi-1 proto-oncogene b	201.04	8.781	2.23E-302	
fli1b	Fli-1 proto-oncogene, ETS transcription factor b	222.47	8.946	0	
dre-mir-142a	dre-mir-142a	119.29	7.857	8.92E-20	NOVEL
mmp13a	matrix metalloproteinase 13a	143.54	8.771	2.13E-108	
flt4	fms-related tyrosine kinase 4	85.28	8.130	8.55E-248	
erg	v-ets avian erythroblastosis virus E26 oncogene homolog	86.16	8.257	2.16E-158	
rasip1	Ras interacting protein 1	137.54	8.959	5.17E-296	
slc29a1b	solute carrier family 29 (equilibrative nucleoside transporter), member 1b	106.65	8.723	9.11E-170	
cltc2	chloride intracellular channel 2	88.71	8.380	1.72E-108	
scarf1	scavenger receptor class F, member 1	103.74	8.740	9.95E-280	
tie1	tyrosine kinase with immunoglobulin-like and EGF-like domains 1	146.15	9.327	3.25E-306	
ncam3	neural cell adhesion molecule 3	93.73	8.740	5.95E-151	
cdh5	cadherin 5	116.26	9.134	0	
lygl2	lysozyme g-like 2	78.10	8.438	5.07E-74	NOVEL
ikzf1	IKAROS family zinc finger 1 (Ikaros)	71.73	8.671	1.27E-122	
tnfrsf2b	tumor necrosis factor, alpha-induced protein 2b	78.70	9.161	1.46E-146	NOVEL
il6r	interleukin 6 receptor	32.42	8.315	4.78E-130	
gata1a	GATA binding protein 1a	50.12	8.167	6.26E-13	
kdr	kinase insert domain receptor (a type III receptor tyrosine kinase)	38.59	9.131	1.19E-157	
myo1f	myosin 1F	26.12	8.544	9.47E-105	
pik3r5	phosphoinositide-3-kinase, regulatory subunit 5	16.95	7.963	1.75E-60	
itgb2	integrin, beta 2	17.18	8.079	5.31E-58	
lpar5a	lysophosphatidic acid receptor 5a	30.36	9.057	3.56E-51	NOVEL
hbae1	hemoglobin, alpha embryonic 1	33.52	8.739	1.12E-20	
ecscr	endothelial cell surface expressed chemotaxis and apoptosis regulator	28.72	9.410	1.05E-92	
sidkey-261j4.3	sidkey-261j4.3	13.25	8.060	1.58E-29	NOVEL
cmklr1	chemokine-like receptor 1	18.75	8.441	7.85E-29	NOVEL
ccr12b.2	chemokine (C-C motif) receptor 12b, tandem duplicate 2	12.20	8.057	6.38E-26	
slc22a7b.1	solute carrier family 22, member 7b, tandem duplicate 1	29.73	9.532	1.28E-48	
coro2bb	coronin, actin binding protein, 2Bb	18.32	8.897	3.10E-33	NOVEL
inpp5d	inositol polyphosphate-5-phosphatase D	7.56	7.899	3.91E-48	
myct1a	myc target 1a	10.78	7.930	8.21E-25	
agtr1b	angiotensin II receptor, type 1b	15.78	8.647	3.31E-33	
si:ch73-208g10.1	si:ch73-208g10.1	10.40	8.503	7.25E-70	
sidkey-119g10.4	sidkey-119g10.4	19.33	8.292	2.28E-16	NOVEL
slc22a7b.2	solute carrier family 22, member 7b, tandem duplicate 2	5.28	7.965	9.16E-27	NOVEL
myct1b	myc target 1b	26.19	9.301	6.14E-24	
si:ch73-90p23.1	si:ch73-90p23.1	10.29	8.096	6.97E-15	NOVEL
fgd5b	FYVE, RhoGEF and PH domain containing 5b	5.41	7.885	4.33E-21	NOVEL
fermt3b	fermitin family member 3b	8.44	9.083	2.24E-30	

Table of the top enriched genes in the 20ss anterior *scf*<sup>+</sup> population upon comparison to expression in the 20ss anterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment. Novel genes have not been previously reported in the literature.

### 20ss anterior depleted genes compared to negative controls

2167 significantly depleted genes in 20ss anterior *scf*<sup>+</sup> cells were associated with 223 over-represented biological GO terms. Depleted tissue-specific processes related to eye and brain development.

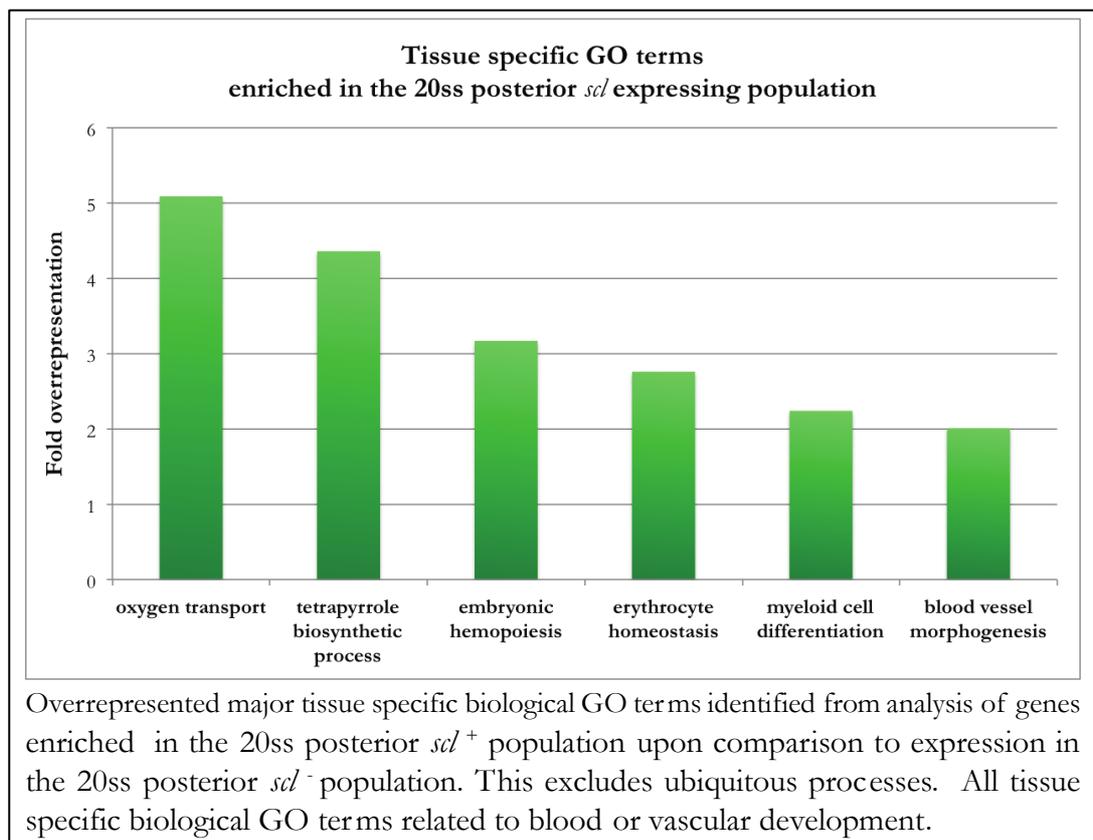
The top over-represented depleted biological GO term was microtubule anchoring, which was over-represented by 12.8 fold in the depleted gene list. The appearance

of this term could suggest that the *scf*<sup>+</sup> cells are one of the most mobile populations of the 20ss anterior. This correlates with the enriched functions of this population in vascular formation and mediating cardiac migration. Genes correlating with ATP synthesis were also over-represented in the list of depleted genes, which could possibly be interpreted as the *scf*<sup>+</sup> cells are less metabolically active than other tissues in the developing anterior of the zebrafish.

#### 4.5.5. 20ss Posterior

5947 genes significantly differentially expressed genes were identified through DESeq2 analysis of the citrine<sup>+</sup> and citrine<sup>-</sup> transcriptomes of the 20ss posterior. As previously annotations from the PANTHER database were used to assess these genes for biological and molecular functions that showed over-representation compared to a randomised background (figure 4-20).

**Figure 4-20 Over-represented tissue-specific gene ontology terms for genes enriched in the 20ss posterior *scf* expressing population.**



## 20ss posterior enriched genes over negative control

4395 genes were significantly enriched in 20ss posterior *sc<sup>+</sup>* cells in comparison to *sc<sup>-</sup>* cells of the 20ss posterior. These enriched genes related to 167 over-represented biological GO categories and included 886 unclassified genes. 18 tissue-specific biological GO terms were over-represented- all of which were haematopoietic or vascular processes. Transcription, translation and tRNA synthesis were also highly over-represented biological processes, which could mean that the *sc<sup>+</sup>* cells of the 20ss are producing proteins at a significantly greater rate than surrounding cells of the posterior. In combination with the over-representation of genes involved in tetrapyrrole biosynthesis, which is the haem production pathway, this enrichment in protein synthesis genes suggests that at this stage the posterior *sc<sup>+</sup>* population produces large amounts of haem.

### Oxygen transport

Oxygen transport is the top biological GO term for the 20ss posterior *sc<sup>+</sup>* population, over-represented by 5.37 fold, correlating with the erythroid lineage being strongly favoured and restricted to the *sc<sup>+</sup>* population.

12 genes contributed to the term oxygen transport- these were six embryonic haemoglobin subunits, their isoforms and myoglobin. Compared to the 20ss anterior and the 10ss posterior, these haemoglobin genes are expressed at dramatically increased levels as detailed in table 4-21. The exceptionally high haemoglobin expression in *sc<sup>+</sup>* cells and subsequent translation may account for some of the over-representation of biological processes involved with protein production. In comparison to the 20ss anterior and 10ss posterior, expression levels of the haemoglobin genes confirm that the main erythroid compartment mainly develops in the posterior of zebrafish embryos and dramatically increases in oxygen transport function between the 10 and 20ss.

**Table 4-21 Genes annotated with the GO term oxygen transporter enriched in the 20ss posterior *scf* expressing population.**

Gene Name	Description	Mean FPKM	log2 FoldChange	p-value
hbbe2	hemoglobin beta embryonic-2	1078.15	6.4255	1.88E-95
HBZ	hemoglobin zeta	215.73	6.2758	9.61E-100
hbz	hemoglobin zeta	123.01	5.9058	1.93E-51
hbbe1.2	hemoglobin beta embryonic-1.2	211.58	5.6805	3.50E-45
hbae1	hemoglobin, alpha embryonic 1	2678.09	4.9574	3.48E-67
hbae3	hemoglobin alpha embryonic-3	7583.52	4.7771	4.84E-76
hbbe3	hemoglobin beta embryonic-3	8086.59	4.6618	2.12E-80
hbae1	hemoglobin, alpha embryonic 1	63.21	4.5813	4.25E-29
hbbe1.1	hemoglobin beta embryonic-1.1	2417.19	4.3936	9.80E-40
hbbe1.1	hemoglobin beta embryonic-1.1	2911.75	3.9544	8.25E-41
hbae1	hemoglobin, alpha embryonic 1	431.08	3.6517	3.75E-30
mb	myoglobin	5.85	2.0385	3.68E-06

Table of oxygen transporter proteins, identified through biological gene ontology analysis of genes enriched in the 20ss posterior *scf*<sup>+</sup> population upon comparison to expression in the 20ss posterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

### DNA replication

DNA replication is required for proliferation of nearly all eukaryotic cell types. Unlike in mammals, zebrafish erythrocytes are nucleated<sup>360</sup>, thus over-representation of biological GO terms relating to DNA replication can be interpreted as an indicator of significant proliferation, including that of the erythroid lineage. In light of the high enrichment of erythroid and haemoglobin genes, I suggest that the erythroid compartment of this posterior *scf*<sup>+</sup> population may undergo rapid proliferation at this stage, potentially in preparation for the commencement of circulation.

### Embryonic haematopoiesis

embryonic haematopoiesis was another over-represented biological GO term for the factors significantly enriched in the 20ss posterior *scf*<sup>+</sup> population and are listed in Table 4-22. These enriched genes are mostly shared with the enriched 20ss anterior haematopoietic factors (many of which have previously been discussed). This includes *lmo2* and *etv2*, which are involved in the development of multiple haemangiogenic lineages. *lmo2* expression is slightly greater in the posterior at the

20ss, however *etv2* is significantly decreased compared to the 20ss anterior. This reduction may signify the differentiation of this *scf*<sup>+</sup> population away from common progenitors into an erythroid population that employs different ETS-family members to complex with Scl and other key transcription factors. Differentiation of this population into a more specifically erythroid population is also suggested by the increase in expression of key erythroid factors, and reduction of vasculogenic expression (*erg*), in comparison to the 20ss anterior *scf*<sup>+</sup> cells.

**Table 4-22 Genes annotated with the GO term embryonic haematopoiesis enriched in the 20ss posterior *scf* expressing population.**

Gene Name	Description	Mean FPKM	log2 FoldChange	p-value
gfi1aa	growth factor independent 1A transcription repressor a	793.09	9.640	1.64E-301
gata1a	GATA binding protein 1a	470.48	7.569	3.96E-115
lmo2	LIM domain only 2 (rhombotin-like 1)	1438.30	7.053	1.65E-191
etv2	ets variant 2	598.63	6.528	2.98E-156
epb41b	erythrocyte membrane protein band 4.1b	75.65	6.336	3.54E-171
tal1	T-cell acute lymphocytic leukemia 1	598.22	6.221	4.15E-145
myct1a	myc target 1a	10.54	6.059	7.05E-25
alas2	aminolevulinate, delta-, synthase 2	352.61	6.013	1.26E-98
sptb	spectrin, beta, erythrocytic	122.65	5.915	8.82E-174
slc25a37	solute carrier family 25, member 37	109.27	5.833	4.14E-102
slc4a1a	solute carrier family 4, member 1a (Diego blood group)	234.69	5.793	4.25E-11
erg	v-ets avian erythroblastosis virus E26 oncogene homolog	16.02	5.715	4.42E-45
csnrp1a	cysteine-serine-rich nuclear protein 1a	119.09	5.191	3.89E-82
cebpa	CCAAT/enhancer binding protein (C/EBP), alpha	168.40	4.414	2.33E-94
trim2a	tripartite motif containing 2a	4.39	3.813	3.85E-16
tmem88a	transmembrane protein 88 a	73.47	3.731	8.32E-51
irak3	interleukin-1 receptor-associated kinase 3	4.42	3.688	2.02E-08
socs1a	suppressor of cytokine signaling 1a	4.87	2.812	5.81E-05
tll1	tolloid-like 1	2.10	2.397	6.94E-09
tspo	translocator protein	112.80	2.390	1.76E-24
snrkb	SNF related kinase b	4.74	1.922	7.10E-06
gata2a	GATA binding protein 2a	14.70	1.469	1.37E-07
dhx8	DEAH (Asp-Glu-Ala-His) box polypeptide 8	24.36	1.259	6.24E-07

Table of factors involved in embryonic haematopoiesis, identified through biological gene ontology analysis of genes enriched in the 20ss posterior *scf*<sup>+</sup> population upon comparison to expression in the 20ss posterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

### Erythroid differentiation

In the 10ss posterior *scf*<sup>+</sup> enriched over negatives gene list erythroid development was a highly over-represented term for the enriched genes. Many of these factors annotated as associated with erythroid development at the 10ss are also included in the erythroid differentiation list of genes enriched within the 20ss posterior *scf*<sup>+</sup>

population (see tables 4-10 and 4-23). However these common genes show a decrease in expression level and are accompanied by the increased enrichment of other erythroid factors such as *gfi1aa*, *slc4a1a* and *epb41b*. This shift in enriched genes could be the result of early factors required to specify the lineage being replaced by factors that contribute later to erythroid functionality, as the population develops from 10ss to 20ss. However both lists contain several haematopoietic factors associated with a range of activities, not limited to the erythroid fate.

**Table 4-23 Genes annotated with the GO term erythroid differentiation enriched in the 20ss posterior *scl* expressing population.**

Gene Name	Description	Mean FPKM	log2 FoldChange	p-value
<i>gfi1aa</i>	growth factor independent 1A transcription repressor a	793.09	9.640	1.64E-301
<i>gata1a</i>	GATA binding protein 1a	470.48	7.569	3.96E-115
<i>spi1b</i>	Spi-1 proto-oncogene b	31.29	7.180	2.01E-65
<i>lmo2</i>	LIM domain only 2 (rhombotin-like 1)	1438.30	7.053	1.65E-191
<i>epb41b</i>	erythrocyte membrane protein band 4.1b	75.65	6.336	3.54E-171
<i>alas2</i>	aminolevulinate, delta-, synthase 2	352.61	6.013	1.26E-98
<i>sptb</i>	spectrin, beta, erythrocytic	122.65	5.915	8.82E-174
<i>slc25a37</i>	solute carrier family 25 (mitochondrial iron transporter), member 37	109.27	5.833	4.14E-102
<i>slc4a1a</i>	solute carrier family 4 (anion exchanger), member 1a (Diego blood group)	234.69	5.793	4.25E-11
<i>fech</i>	ferrochelatase	122.42	5.151	2.34E-79
<i>tspo</i>	translocator protein	112.80	2.390	1.76E-24
<i>slc25a38b</i>	solute carrier family 25, member 38b	8.38	1.768	1.92E-11
<i>tmod2</i>	tropomodulin 2	25.59	1.698	7.63E-11
<i>maca</i>	macrophage erythroblast attacher	29.23	1.609	1.80E-11
<i>pcdcd2</i>	programmed cell death 2	34.32	1.590	7.10E-08
<i>rasa3</i>	RAS p21 protein activator 3	11.62	1.585	8.75E-09
<i>gata2a</i>	GATA binding protein 2a	14.70	1.469	1.37E-07
<i>snx3</i>	sorting nexin 3	37.67	1.272	1.15E-05
<i>melk</i>	maternal embryonic leucine zipper kinase	15.42	1.115	1.09E-04
<i>vhl</i>	von Hippel-Lindau tumor suppressor	8.96	0.867	1.97E-02
<i>jak2b</i>	Janus kinase 2b	8.74	0.855	1.69E-03
<i>tert</i>	telomerase reverse transcriptase	10.27	0.765	2.68E-03
<i>cdc73</i>	cell division cycle 73, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)	45.92	0.748	6.93E-04
<i>atp1f1b</i>	ATPase inhibitory factor 1b	123.56	0.493	2.98E-02

Table of factors involved in erythroid differentiation, identified through biological gene ontology analysis of genes enriched in the 20ss posterior *scl*<sup>+</sup> population upon comparison to expression in the 20ss posterior *scl*<sup>-</sup> population. Factors are in order of fold enrichment.

### Blood vessel morphogenesis

Blood vessel morphogenesis was an over-represented biological GO term for both the 10 and 20ss posterior *scl*<sup>+</sup> populations. The number of genes associated with this term in the 20ss is greater and includes additional factors such as *e2f7*, *e3f8*, and *sars* (table 4-24). The transcription factor genes *e2f7* and *e2f8* are capable of binding and promoting expression of VEGF during sprouting angiogenesis to stimulate and guide blood vessel development<sup>361</sup>.

**Table 4-24 Genes annotated with the GO term blood vessel morphogenesis enriched in the 20ss posterior *scf* expressing population.**

Gene Name	Description	Mean FPKM	log2 FoldChange	p-value
<i>fli1b</i>	Fli-1 proto-oncogene, ETS transcription factor b	69.70	7.071	1.75E-131
<i>lmo2</i>	LIM domain only 2 (rhombotin-like 1)	1438.30	7.053	1.65E-191
<i>ecscr</i>	endothelial cell surface expressed chemotaxis and apoptosis regulator	15.21	6.994	1.94E-71
<i>sox7</i>	SRY (sex determining region Y)-box 7	309.74	6.937	8.54E-165
<i>etv2</i>	ets variant 2	598.63	6.528	2.98E-156
<i>cdh5</i>	cadherin 5	37.40	6.461	1.45E-158
<i>kdr</i>	kinase insert domain receptor (a type III receptor tyrosine kinase)	9.93	6.162	1.17E-08
<i>fli1a</i>	Fli-1 proto-oncogene, ETS transcription factor a	281.84	5.937	4.61E-118
<i>erg</i>	v-ets avian erythroblastosis virus E26 oncogene homolog	16.02	5.715	4.42E-45
<i>flt1</i>	fms-related tyrosine kinase 1 (vascular endothelial growth factor receptor)	3.15	5.137	1.30E-24
<i>si:ch73-334d15.2</i>	<i>si:ch73-334d15.2</i>	5.85	5.047	8.03E-15
<i>arhgef9b</i>	Cdc42 guanine nucleotide exchange factor (GEF) 9b	12.24	5.042	4.37E-63
<i>sox18</i>	SRY (sex determining region Y)-box 18	34.92	5.014	2.58E-49
<i>hspa12b</i>	heat shock protein 12B	10.28	4.495	3.58E-39
<i>dab2</i>	Dab, mitogen-responsive phosphoprotein, homolog 2 (Drosophila)	29.52	4.435	3.64E-43
<i>hapln1b</i>	hyaluronan and proteoglycan link protein 1b	12.24	4.112	1.93E-35
<i>snx5</i>	sorting nexin 5	178.64	3.327	1.10E-37
<i>plekhg5a</i>	pleckstrin homology domain containing, family G (with RhoGef domain) member 5a	9.11	3.316	3.71E-32
<i>mcamb</i>	melanoma cell adhesion molecule b	45.53	3.293	1.08E-33
<i>nrp1b</i>	neuropilin 1b	17.47	2.845	7.96E-21
<i>ramp2</i>	receptor (G protein-coupled) activity modifying protein 2	43.39	2.794	3.88E-28
<i>dll4</i>	delta-like 4 (Drosophila)	23.21	2.350	4.90E-14
<i>mb</i>	myoglobin	5.85	2.039	3.68E-06
<i>msna</i>	moesin a	219.33	2.015	1.05E-21
<i>fmnl3</i>	formin-like 3	16.99	2.008	5.69E-19
<i>itga5</i>	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	41.05	1.996	1.70E-17
<i>e2f8</i>	E2F transcription factor 8	14.54	1.991	1.57E-15
<i>hey2</i>	hes-related family bHLH transcription factor with YRPW motif 2	69.28	1.915	3.61E-17
<i>rab5c</i>	RAB5C, member RAS oncogene family	67.77	1.631	3.53E-10
<i>arhgef7b</i>	Rho guanine nucleotide exchange factor (GEF) 7b	26.06	1.604	1.54E-12
<i>gdf6a</i>	growth differentiation factor 6a	30.98	1.599	2.84E-10
<i>lpar2a</i>	lysophosphatidic acid receptor 2a	24.64	1.482	1.85E-07
<i>gata2a</i>	GATA binding protein 2a	14.70	1.469	1.37E-07
<i>wdr43</i>	WD repeat domain 43	236.01	1.439	1.42E-09
<i>sat1b</i>	spermidine/spermine N1-acetyltransferase 1b	8.17	1.290	5.89E-04
<i>sars</i>	seryl-tRNA synthetase	74.81	1.073	3.35E-06
<i>sh3gl3b</i>	SH3-domain GRB2-like 3b	7.37	1.066	7.57E-04
<i>smad5</i>	SMAD family member 5	75.49	1.059	6.99E-06
<i>cdc42</i>	cell division cycle 42	121.62	1.056	1.66E-06
<i>aggf1</i>	angiogenic factor with G patch and FHA domains 1	16.42	0.994	5.68E-05
<i>colec12</i>	collectin sub-family member 12	27.58	0.977	6.57E-05
<i>rab11a</i>	RAB11a, member RAS oncogene family	43.01	0.972	9.84E-06
<i>vegfc</i>	vascular endothelial growth factor c	5.60	0.945	4.79E-03
<i>ell</i>	elongation factor RNA polymerase II	12.81	0.914	1.34E-04
<i>gna13b</i>	guanine nucleotide binding protein (G protein), alpha 13b	14.30	0.904	3.60E-03
<i>e2f7</i>	E2F transcription factor 7	10.46	0.887	7.79E-03
<i>ppp4ca</i>	protein phosphatase 4, catalytic subunit a	16.43	0.823	1.45E-03
<i>ppp1cab</i>	protein phosphatase 1, catalytic subunit, alpha isozyme b	114.90	0.794	1.80E-04
<i>eif3i</i>	eukaryotic translation initiation factor 3, subunit I	195.08	0.754	2.93E-04
<i>gng2</i>	guanine nucleotide binding protein (G protein), gamma 2	45.63	0.723	3.43E-03
<i>rtn4a</i>	reticulon 4a	106.69	0.713	3.67E-03
<i>stk25b</i>	serine/threonine kinase 25b	28.84	0.701	2.20E-02
<i>lama4</i>	laminin, alpha 4	10.88	0.684	4.25E-03
<i>shc1</i>	SHC (Src homology 2 domain containing) transforming protein 1	15.63	0.678	6.51E-03
<i>pik3c2a</i>	phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 alpha	3.92	0.657	2.25E-02
<i>kif11</i>	kinesin family member 11	42.60	0.567	1.14E-02
<i>cds2</i>	CDP-diacylglycerol synthase (phosphatidate cytidyltransferase) 2	26.99	0.561	1.57E-02
<i>gnb2l1</i>	guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1	1824.71	0.481	1.79E-02

Table of factors involved in blood vessel morphogenesis, identified through biological gene ontology analysis of genes enriched in the 20ss posterior *scf*<sup>+</sup> population upon comparison to expression in the 20ss posterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

*sars* is a tRNA synthetase that has been investigated for its role in angiogenic sprouting<sup>362</sup> following the identification of multiple mutants that showed defective blood vessel organisation and related to the *sars* gene.

Most of the commonly expressed factors show a decrease in expression as this population develops from the 10 to 20ss. This decrease in expression of early expressed factors and increase in genes associated with vessel sprouting and tissue infiltration could suggest that by 20ss the posterior *sc<sup>l</sup>* population has formed the endothelial cells of the main vessels and the remaining vascular function relates to capillary sprouting and addition to existing vessels.

### **Molecular function of 20ss posterior enriched genes over negative controls**

As in the 10ss posterior *sc<sup>l</sup>* population oxygen transport was the top over-represented molecular function following analysis using the PANTHER database. Also similarly to the 10ss posterior, histone methyl-transferase activity was also a highly over-represented molecular GO term for the enriched genes of the 10ss posterior *sc<sup>l</sup>* population. However at the 10ss this term was specifically histone lysine methylation, at the 20ss both lysine and arginine histone methylases were featured in this over-represented group. Table 4-25 details the significantly enriched genes that are annotated as histone methyltransferases of the posterior at the 20ss. The histone arginine methyltransferases enriched in the 20ss posterior *sc<sup>l</sup>* cells, include transcriptionally activating orthologues of *prmt1* and *carm1* and transcriptionally repressive *prmt5* and 7. The enriched expression of this histone modifying enzymes within the developing erythroid compartment suggests that, in comparison to surrounding tissues, epigenetic regulation of transcription is more prominent in posterior *sc<sup>l</sup>* expressing populations.

**Table 4-25 Histone methyl transferase genes enriched in the 20ss posterior *scf* expressing population.**

Gene Name	Description	Mean FPKM	log2 FoldChange	p-value
suv39h1b	suppressor of variegation 3-9 homolog 1b	48.38	2.422	4.71E-25
men1	multiple endocrine neoplasia I	57.21	2.055	1.27E-17
setdb1a	SET domain, bifurcated 1a	41.09	2.047	3.44E-17
suv420h2	suppressor of variegation 4-20 homolog 2 (Drosophila)	16.76	1.683	3.07E-11
prmt7	protein arginine methyltransferase 7	61.29	1.675	2.34E-08
prmt3	protein arginine methyltransferase 3	44.80	1.548	1.69E-10
prdm9	PR domain containing 9	13.38	1.451	1.57E-08
setd2	SET domain containing 2	8.73	1.045	6.40E-05
prmt1	protein arginine methyltransferase 1	798.73	1.009	1.10E-06
dnmt1	DNA (cytosine-5-)-methyltransferase 1	31.23	0.911	1.26E-04
carm1	coactivator-associated arginine methyltransferase 1	112.53	0.909	3.11E-04
nsd1a	nuclear receptor binding SET domain protein 1a	14.72	0.902	1.67E-04
kmt2ca	lysine (K)-specific methyltransferase 2Ca	3.33	0.900	7.92E-04
prmt5	protein arginine methyltransferase 5	73.19	0.899	3.10E-04
suv420h1	suppressor of variegation 4-20 homolog 1 (Drosophila)	25.97	0.856	1.22E-03
kmt2d	lysine (K)-specific methyltransferase 2D	4.69	0.838	2.95E-03
kmt2ba	lysine (K)-specific methyltransferase 2Ba	5.58	0.815	5.45E-03
dot1l	DOT1-like histone H3K79 methyltransferase	5.80	0.803	5.17E-03
kmt2cb	lysine (K)-specific methyltransferase 2Cb	6.34	0.692	1.60E-02
whsc1	Wolf-Hirschhorn syndrome candidate 1	21.47	0.592	1.31E-02
ash2l	ash2 (absent, small, or homeotic)-like (Drosophila)	56.22	0.542	2.08E-02

Table of histone methyl-transferase genes enriched in the 20ss posterior *scf*<sup>+</sup> population upon comparison to expression in the 20ss posterior *scf*<sup>-</sup> population. Factors are in order of fold enrichment.

### Protein class analysis of 20ss posterior enriched genes over negative controls

Transcription factors represented 7.5% of enriched genes in the 20ss posterior *scf*<sup>+</sup> cells as annotated by the PANTHER database and 2.9% of genes were labelled as signalling molecules. Highly over-represented biological GO term for the transcription factor list included primitive haematopoiesis, myeloid cell differentiation and blood vessel morphogenesis. Surprisingly no erythroid related terms were indicated as being significantly over-represented. GO analysis of the total enriched genes clearly shows that this population has a strongly erythroid profile. However genes annotated with a primitive haematopoiesis biological term include key erythroid factors (*gata1*, *gata2* and *gfi1a*). This absence of erythroid terms is also observed in the 10ss posterior enriched transcription factor analyses. The lack of erythroid GO classes for enriched transcription factors in both of these strongly

erythroid cell types suggested that this is an artefact of annotation rather than a biological feature.

### **Top enriched gene in the 20ss posterior, compared to negative controls**

All but one of the top 50 enriched genes can be used to calculate rational enrichment values, as they had mapped transcripts in the *sc<sup>l</sup>-* sample, and are listed in table 4-26. 19 of these top enriched genes have been previously shown to contribute to haematopoietic or vascular development, 10 genes have published expression within the developing circulatory system of zebrafish or seen to be involved in haemangiogenic processes in other models. 18 factors highly enriched in the 20ss *sc<sup>l</sup><sup>+</sup>* posterior population were identified as novel.

### **20ss posterior depleted genes compared to negative controls**

163 biological terms were determined by analysis using the PANTHER database annotations to be significantly depleted in the *sc<sup>l</sup><sup>+</sup>* cells of the 20ss posterior compared to their *sc<sup>l</sup><sup>-</sup>* neighbours. Genes for neuron migration were highly over-represented in the depleted gene list. Interestingly blood vessel development also appeared as a biological GO term that was over-represented in the depleted genes of the 20ss posterior *sc<sup>l</sup><sup>+</sup>* population. Wnt and Notch signalling pathways were also featured in these GO terms reduced in *sc<sup>l</sup><sup>+</sup>* cells. These signalling cascades are crucial for earlier stages of haematopoietic development<sup>241,356,363,364</sup>, though depletion of their factors in 20ss posterior *sc<sup>l</sup><sup>+</sup>* cells could suggest that these signalling pathways are no longer required as erythroid development proceeds. Muscle, kidney, somite, brain and eye development biological GO terms were also observed to be depleted in the *sc<sup>l</sup><sup>+</sup>* cells of the 20ss posterior.

**Table 4-26 Top enriched genes in the 20ss posterior *sc*<sup>+</sup> expressing population**

Gene Name	Description	Mean FPKM	log2 FoldChange	p-value	Novel?
gfi1aa	growth factor independent 1A transcription repressor a	793.09	9.640	1.64E-301	
sele	selectin E	6.88	9.438	7.31E-30	
blf	bloody fingers	1567.13	8.246	8.89E-249	
mpx	myeloid-specific peroxidase	8.15	7.946	2.64E-34	
tnfaip2b	tumor necrosis factor, alpha-induced protein 2b	35.51	7.936	5.18E-85	NOVEL
f2r	coagulation factor II (thrombin) receptor	161.69	7.873	2.06E-124	
si:ch73-248e21.5	si:ch73-248e21.5	6.04	7.817	4.02E-28	NOVEL
ikzf1	IKAROS family zinc finger 1 (Ikaros)	106.80	7.745	4.17E-163	
morc3b	MORC family CW-type zinc finger 3b	202.44	7.594	8.22E-08	
lygl2	lysozyme g-like 2	75.06	7.593	1.05E-76	NOVEL
gata1a	GATA binding protein 1a	470.48	7.569	3.96E-115	
aqp1a.2	aquaporin 1a (Colton blood group), tandem duplicate 2	3.49	7.558	2.02E-08	
si:dkey-183p4.9	si:dkey-183p4.9	7.39	7.458	2.76E-12	NOVEL
TRAF3IP3	TRAF3 interacting protein 3	2.23	7.427	3.55E-16	NOVEL
lrrc15	leucine rich repeat containing 15	32.86	7.335	1.13E-91	
klf17	Kruppel-like factor 17	1293.56	7.194	3.44E-185	
spi1b	Spi-1 proto-oncogene b	31.29	7.180	2.01E-65	
si:ch73-299h12.1	si:ch73-299h12.1	593.97	7.170	1.35E-129	NOVEL
rfsd	Rieske (Fe-S) domain containing	811.29	7.170	1.85E-134	NOVEL
PARM1	prostate androgen-regulated mucin-like protein 1	24.16	7.110	2.31E-47	NOVEL
SUSD1	si:ch73-247j11.2	73.61	7.075	1.70E-151	NOVEL
fli1b	Fli-1 proto-oncogene, ETS transcription factor b	69.70	7.071	1.75E-131	
lmo2	LIM domain only 2 (rhombotin-like 1)	1438.30	7.053	1.65E-191	
dre-mir-142a	dre-mir-142a	110.75	7.011	5.53E-20	NOVEL
si:dkey-261j4.4	si:dkey-261j4.4	18.80	7.001	3.08E-64	NOVEL
ecscr	endothelial cell surface expressed chemotaxis and apoptosis regulator	15.21	6.994	1.94E-71	
col8a2	collagen, type VIII, alpha 2	30.85	6.992	4.93E-74	
aqp8a.1	aquaporin 8a, tandem duplicate 1	110.75	6.963	3.31E-121	
sox7	SRY (sex determining region Y)-box 7	309.74	6.937	8.54E-165	
gfi1b	growth factor independent 1B transcription repressor	249.67	6.871	3.92E-172	
PAQR9	progesterin and adipoQ receptor family member IX	20.42	6.858	7.53E-42	NOVEL
si:ch73-248e21.7	si:ch73-248e21.7	141.74	6.857	1.39E-163	NOVEL
drl	draculin	103.16	6.854	2.63E-132	
cldng	claudin g	146.41	6.827	1.23E-123	
rasip1	Ras interacting protein 1	53.27	6.791	7.48E-111	
agtr2	angiotensin II receptor, type 2	53.73	6.754	1.04E-73	
cpx	coproporphyrinogen oxidase	1010.97	6.731	1.48E-118	
zgc:171686	zgc:171686	43.74	6.726	3.25E-75	NOVEL
slc29a1b	solute carrier family 29 (equilibrative nucleoside transporter), member 1b	16.00	6.723	4.39E-65	
tie1	tyrosine kinase with immunoglobulin-like and EGF-like domains 1	41.72	6.706	5.06E-124	
aqp8a.2	aquaporin 8a, tandem duplicate 2	4.05	6.704	1.30E-12	
gmip	GEM interacting protein	6.59	6.701	2.54E-49	NOVEL
slc22a7b.1	solute carrier family 22, member 7b, tandem duplicate 1	13.91	6.647	2.85E-45	
ela2	elastase 2	24.59	6.631	2.89E-41	
CR391944.2	Novel	19.33	6.625	3.51E-66	NOVEL
fgf21	fibroblast growth factor 21	8.96	6.609	1.23E-14	
si:dkey-261j4.3	si:dkey-261j4.3	59.44	6.595	9.31E-82	NOVEL
si:ch73-147f11.1	si:ch73-147f11.1	149.23	6.549	1.67E-152	
etv2	ets variant 2	598.63	6.528	2.98E-156	
myct1b	myc target 1b	8.56	6.524	1.44E-17	

Table of the top 50 enriched genes in the 20ss posterior *sc*<sup>+</sup> population upon comparison to expression in the 20ss posterior *sc*<sup>-</sup> population. Factors are in order of fold enrichment. Novel genes have not been previously reported in the literature.

## 4.6. Chapter 4 summary

This project aimed to investigate the role of *sc<sup>+</sup>* cells in haematopoietic and vascular development. I have generated genome wide datasets for transcription in four *sc<sup>+</sup>* *in vivo* contexts. Using the transgenic line that I have generated I have specifically isolated *sc<sup>+</sup>* populations from the anterior and posterior of 10 and 20ss zebrafish embryos. From these samples RNA was isolated and poly-A selected, the resulting RNA was sequenced and I have mapped to the zebrafish genome. Transcriptomic datasets were highly reproducible and identified distinct profiles for each of the four *sc<sup>+</sup>* cellular contexts studied.

Comparison of my enriched gene lists to the PANTHER database enabled me to annotate my data with gene ontology (GO) terms. Despite a range of limitations associated with the tool, PANTHER analysis identified significantly over representation of specific gene classes associated with each of the *sc<sup>+</sup>* gene sets. GO analysis was useful to handle such large datasets, but is dependant on the quality of the database annotations and also does not take into account relative expression levels.

Spatial comparison of the 10ss *sc<sup>+</sup>* RNA-seq datasets shows that at this stage the anterior *sc<sup>+</sup>* transcriptome shows a greater diversity of associated biological functions than the posterior *sc<sup>+</sup>* population. Conversely this early 10ss posterior population shows a strongly erythroid profile, including high expression of haemoglobin genes.

Temporal comparison of the posterior *sc<sup>+</sup>* population identifies a strong erythroid profile that is present at the 10ss but augmented and increases in expression level in the 20ss. The 20ss posterior population also displays significantly higher expression than the 10ss posterior of certain key genes that have previously been associated with myeloid development.

Over 8500 genes were identified as expressed in all four *sc<sup>l</sup><sup>+</sup>* contexts, 5000 of which showed a common expression level across the *sc<sup>l</sup><sup>+</sup>* cellular contexts. Such commonly expressed genes may represent a core *sc<sup>l</sup>* regulatory network that functions across a range of biological contexts. Supporting this theory the only tissue-specific GO terms over-represented by these commonly expressed shared genes are related to haematopoiesis. These commonly expressed shared genes may be an ideal starting point to investigate novel factors that may be involved in early haematopoietic and vascular developmental programs. To confirm the presence of a core *sc<sup>l</sup>* regulatory network active in these contexts several further investigations would be required.

These include:

- Confirmation of co-expression of factors within single *sc<sup>l</sup><sup>+</sup>* cells of each context.
- Evidence that co-expressed factors mediate similar biological functions in each of the studied contexts.
- Evidence that co-expressed factors interact within a single regulatory network that is shared by all four contexts.

To further resolve the specific transcriptomes associated with each early *sc<sup>l</sup><sup>+</sup>* haematopoietic population, enriched transcriptomes for each of the context studied were produced through comparison to neighbouring *sc<sup>l</sup>* populations. Differential expression analysis of the *sc<sup>l</sup><sup>+</sup>* versus the *sc<sup>l</sup><sup>-</sup>* transcriptomes for each of these contexts produced a set of four *sc<sup>l</sup><sup>+</sup>* enriched gene lists. Despite the reliance on published studies for annotation, PANTHER analysis identified significantly over representation of specific gene classes associated with each of the *sc<sup>l</sup><sup>+</sup>* enriched gene sets. This approach does not prove that these activities are occurring within the studied population, for this detection of end products would be required. However this analysis provides a useful overview and confirms that the *sc<sup>l</sup><sup>+</sup>* populations show

statistically significantly different transcriptional profiles compared to their neighbouring *scf*<sup>-</sup> cells.

Many of the enriched factors have been shown to be key regulators and signalling molecules in previous studies using *in situ* techniques, morpholino and over-expression assays in zebrafish, mice and cell culture models.

The 10ss anterior *scf*<sup>+</sup> enriched transcriptome displayed over-representation of genes related to rho signalling, angiogenesis, myelopoiesis and heart development. Key vasculogenic transcription factors were identified as significantly enriched in the 10ss anterior *scf*<sup>+</sup> population, suggesting that blood vessel development is strongly favoured in this context. Enriched factors in the 20ss anterior population show over-representation of only three biological GO terms: angiogenesis, oxygen transport and embryonic haematopoiesis. This analysis suggests the anterior *scf*<sup>+</sup> populations contribute to the development of myelopoietic and erythrogenic lineages in addition to the heart field. This range in biological functions enriched in the anterior *scf*<sup>+</sup> transcriptome correlates with the *in vivo* imaging data of the anterior *citrine*<sup>+</sup> population (see Figure 3-14).

The 10 and 20ss posterior *scf*<sup>+</sup> enriched transcriptomes show enriched expression of factors required for haem synthesis and the cell cycle. Myeloid and vascular factors are also specifically enriched in the transcriptomes of both posterior contexts but at significantly lower levels than genes involved in erythrogenic functions. This analysis suggests that the posterior population is rapid proliferating and mainly contributing to the erythroid lineage, yet there is also some smaller contribution to vascular and myeloid cell fates.

Another key strength of using GO analysis tool in combination with differential expression analysis with genome-wide datasets is the ability to easily identify novel factors that may be interesting for further investigation (For example see table 4-22).

In this chapter I identify highly expressed and significantly enriched factors in each of the  $scf^+$  contexts that currently lack annotation of their biological function. Further investigation into the expression pattern and the biological contribution to haematopoietic development of these novel factors may identify key regulatory genes of interest.

Through this analysis I have determined that the  $scf^+$  transcriptional programs of the anterior and posterior already show distinct enrichment patterns at the 10ss. I have also identified that the 10ss posterior shows a strongly erythrogenic posterior transcriptional program that is enriched compared to  $scf$  neighbouring cells. The 20ss posterior  $scf^+$  population shows a similar enrichment of transcripts for genes associated with key erythrogenic functions, with very high expression of these factors specifically within this  $scf^+$  context.

In conclusion this RNA-seq analysis of these early  $scf$  expressing populations has confirmed co-expression of key factors within a single cell population, *in vivo* while also describing the full transcriptome of each of these populations, including previously unknown factors.

## 5. Investigating early variation in chromatin accessibility within *in vivo* *sc<sup>+</sup>* populations.

### 5.1. Introduction.

Each of these four *sc<sup>+</sup>* contexts shows significantly different transcriptional profiles. Differences in transcriptional output arise from differences in underlying regulatory mechanisms including regulation of transcription and transcript degradation. As an initial investigation into the underlying regulatory programs active in early haematopoiesis I have performed ATAC-seq on *sc<sup>+</sup>* samples for each of these contexts with at least two replicates per context. Details of the read, mapping and peak number are shown in figure 5-1.

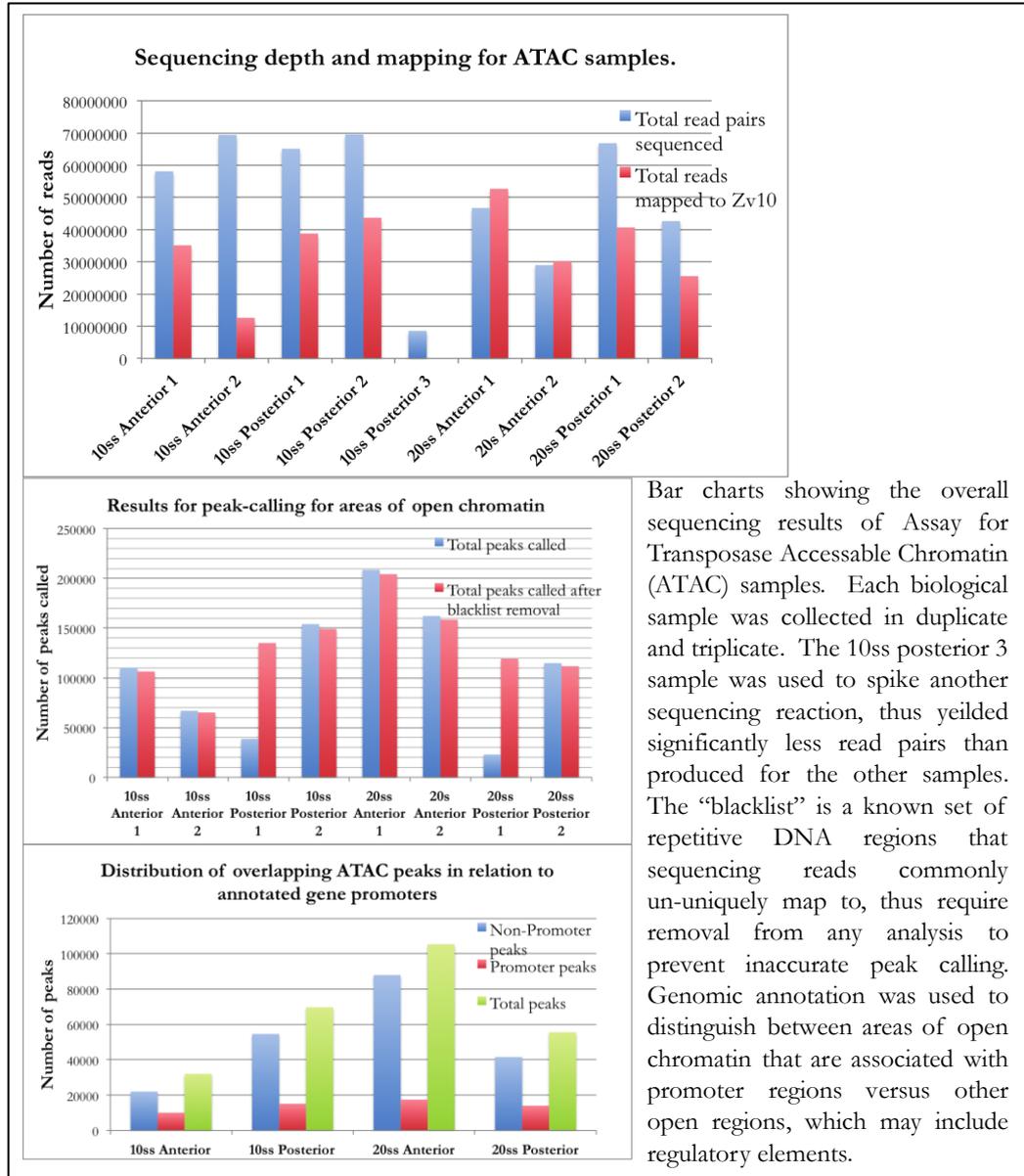
ATAC-seq identifies on a genome-wide scale regions of accessible chromatin. Chromatin becomes accessible to the Tn5 transposase used in this approach during a range of biological processes that require dynamic protein binding to the genome, such as transcription and regulation through enhancer-promoter interactions. Analyses of these genome-wide chromatin accessibility datasets are poised to inform us on regulatory architecture and chromatin landscapes of *sc<sup>+</sup>* cells from different developmental contexts, and permit identification putative regulatory elements within them.

ATAC-seq peaks were called and context-specific datasets were generated. We initially used a very stringent definition of peaks as those present in both replicates. Each ATAC-seq dataset contained a number of promoter-associated peaks, whereas the remaining peaks, which possibly represent putative *cis*-regulatory modules and CTCF-bound regions were designated as non-promoter ATAC-seq peaks.

To generate context-specific ATAC-seq sets of non-promoter genomic regions (putative regulatory elements), we first eliminated ATAC-seq peaks called at promoter from each datasets. Following this we performed comparative analysis to

identify subsets of putative regulatory elements that are uniquely called in one of the two compared developmental contexts.

**Figure 5-1 Overview of ATAC sequencing and mapping for early *scf*<sup>+</sup> cell populations**



## 5.2. Initial differences in chromatin accessibility in early *scf*<sup>+</sup> cell populations.

Table 5-1 shows that we recover comparable number of promoter-associated ATAC-seq peaks for each context. Open promoter loci datasets were correlated with gene expression datasets within the corresponding context (as previously defined with the FPKM value > 2). We observe that anterior *scf*<sup>+</sup> cells at early stage (10ss) have a significant number of loci (~40%) with open promoters, but have not yet

engaged in transcription. This is in contrast to the same population at later stage (20ss), when ~90% of open promoters also produce detectable transcripts. This could suggest that the regulatory program in the anterior *scf*<sup>+</sup> cells onsets around the 10ss and is fully engaged by 20ss. Interestingly the number of open promoters in the anterior *scf*<sup>+</sup> cells does not change drastically between 10ss and 20ss, suggesting that mechanisms for priming and pioneering the system may take place very early during embryogenesis in this lineage.

**Table 5-1 Distribution of promoter associated ATAC-seq peaks, between the four contexts studied.**

Context	Genes with promoter ATAC peaks	Genes expressed	Genes with common peaks
10ss anterior	10216	11112	6879
10ss posterior	14976	10903	9028
20ss anterior	16929	11143	9888
20ss posterior	14000	10197	8407

This interpretation of a primed but not fully engaged anterior *scf* programme is further supported by a matching ‘deficit’ in non-promoter, (i.e. putative cis-regulatory module) ATAC peaks (table 5-2). The number of these accessible non-promoter regions also drastically increases between 10ss and 20ss in the anterior *scf*<sup>+</sup> cells, correlating with an increase in transcription from open loci. Such dynamic activity in opening putative regulatory regions just prior to activation of transcription, has been proven to be an excellent predictor of spatiotemporal specific activity of the developmental *cis*-regulatory modules (personal correspondence Daria Gavriouchkina, not published). This suggests that comparative analysis of our datasets could be used for identification and foot-printing of putative context-specific enhancers, which following validation could be used for the assembly of gene regulatory circuits.

**Table 5-2 Distribution of non-promoter associated ATAC-seq peaks, between the four *sc<sup>l</sup>* contexts studied.**

Context comparison	Non-promoter ATAC peaks	Genes in proximity to non-promoter ATAC peaks
<b>10ss, anterior vs posterior</b>	6851	5203
<b>10ss, posterior vs anterior</b>	39608	14779
<b>20ss, anterior vs posterior</b>	57666	16321
<b>20ss, posterior vs anterior</b>	11103	7388

Conversely, the regulatory dynamics of the posterior *sc<sup>l</sup>* programme is less variable between these developmental time points. Early posterior *sc<sup>l</sup>* cells showed a greater number of promoter and non-promoter open chromatin peaks, compared to the anterior *sc<sup>l</sup>* population at the same stage (figure 5-1). This could be interpreted as the chromatin of the early posterior *sc<sup>l</sup>* cells showing greater accessibility, suggestive of active remodelling, transcription factor binding, and hence more engaged gene regulatory programs than in the comparable anterior population.

This higher early activity is accompanied by the majority of open promoters being associated with transcription in 10ss *sc<sup>l</sup>* posterior population and maintained in the 20ss *sc<sup>l</sup>* posterior population. The numbers of non-promoter peaks (putative open and active regulatory elements) showed a downward trend in number as the posterior *sc<sup>l</sup>* program progressed from the 10ss to 20ss (Table 5-2). This could possibly suggest a reduction or refinement of the regulatory programs active in the posterior *sc<sup>l</sup>* population as the cells progress from the 10 to 20ss.

### 5.3. Analysis of regulatory loci associated with context specific-ATAC peaks

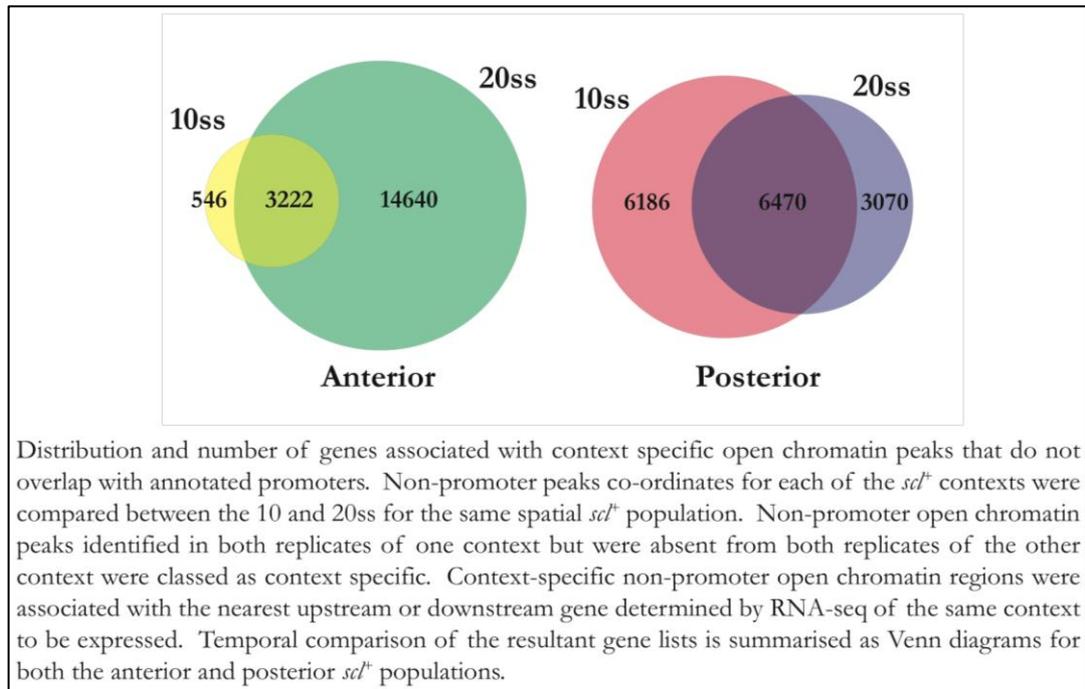
For this experiment, the genomic locations of replicable non-promoter peaks were compared, for either the anterior or posterior *sc<sup>l</sup>* population, between the 10ss and 20ss, or for either the 10ss and 20ss *sc<sup>l</sup>* population, between the anterior or posterior. Context-specific peaks (i.e. non-promoter ATAC-seq peaks found in one context, but not the other context) were associated with the closest expressed genes,

as the most likely candidates those putative elements could be regulating. Expressed genes were based on RNA-seq results for the appropriate context. By focussing on context specific non-promoter peaks, this analysis focuses on regions of dynamically open chromatin, which is more likely to be involved in dynamic cellular processes than ubiquitous or housekeeping activities.

### 5.3.1. Temporal comparison of genes associated with non-promoter ATAC peaks

We analysed the nature and number of genes associated with non-promoter ATAC-seq peaks identified in the same spatial contexts at different developmental time points (See figure 5-2). The comparison of these gene sets offers a preliminary insight into a relative regulatory activity in *sc<sup>fl</sup>* cells both at different embryonic sites (anterior and posterior), but also across developmental time.

**Figure 5-2 Distribution and overlap of non-promoter ATAC peaks in temporal comparisons of *sc<sup>fl</sup>* cellular contexts.**



We observed that the anterior *sc<sup>fl</sup>* population at 10ss has a significantly smaller number of genes (546) associated with non-promoter open chromatin peaks unique to this context, when compared to 20ss, where the number of context-specific

associated open loci is significantly higher (14640). This suggests that in the anterior between the 10 and 20ss chromatin becomes more open at sites away from promoters that may represent regulatory regions. If these non-promoter regions that become opened during the anterior development do correlate with regulatory elements this analysis would suggest that a regulatory program becomes dynamically activated during this time. Through investigation of the *in vivo* activity associated with these putative regulatory elements and confirmation of their target genes it may be possible to suggest that this represents the onset of the anterior haematopoietic lineage specification program.

Conversely, posterior *scf*<sup>+</sup> population at 10ss has a significantly higher number (6168) of genes associated with non-promoter open chromatin regions that are stage specific versus the 20ss sample in this posterior temporal comparison (3070). This analysis shows that by the 10ss in the posterior non-promoter areas of open chromatin are already present and that posterior regulatory activity within *scf*<sup>+</sup> cells, is already underway at 10ss.

Approximately half of the genes associated with non-promoter context specific peaks in the 10ss are also associated with non-promoter context specific peaks in the 20ss posterior sample. This suggests that while the same transcriptional output may be observed the regulatory activity that drives it might be shifting between programs. For further insight into the regulatory activity occurring in the posterior *scf*<sup>+</sup> population, validation of *in vivo* enhancer activity and target gene confirmation would be required. This putative regulatory program shift could represent the switch between a progenitor program and the onset of lineage specification and differentiation.

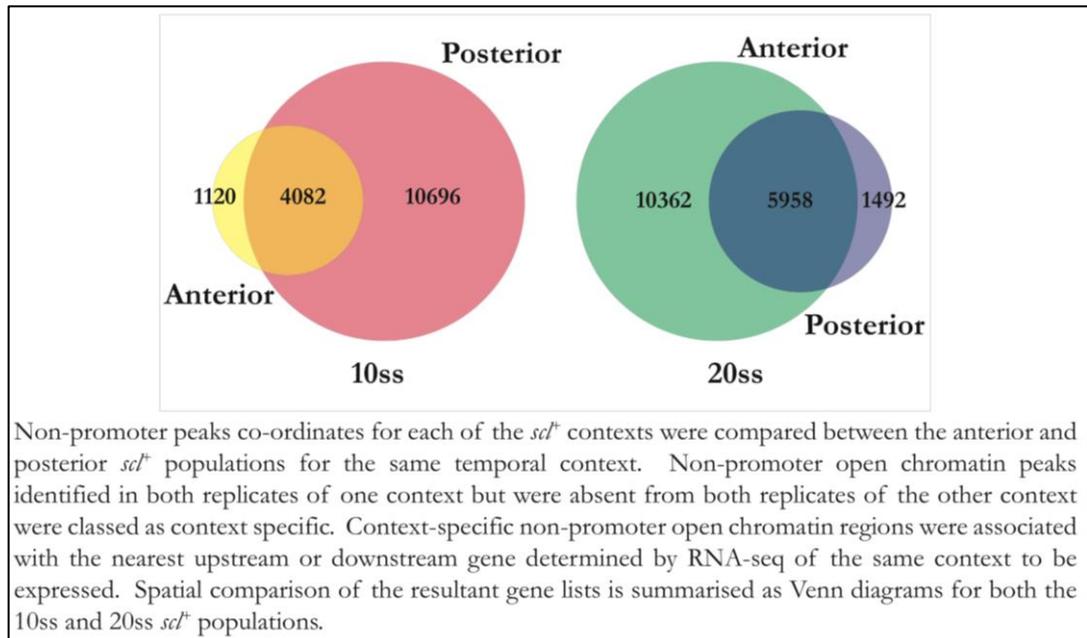
In conclusion the comparison of non-promoter open chromatin peaks for either the anterior or the posterior *scf*<sup>+</sup> population across primitive haematopoietic

development, suggests that the underlying regulatory activity is very different between the two populations. In the anterior this data suggests the onset of regulatory activity driving the opening of chromatin between the 10 and 20ss. In the posterior this data suggests that an initial regulatory program is already active and being superseded by a second overlapping program, during these time points. However this is very preliminary data that can be used to direct further investigations but requires further validation before any firm conclusions can be drawn.

### 5.3.2. Spatial comparison of genes associated with non-promoter ATAC-peaks

Upon spatial comparison of non-promoter peaks of chromatin accessibility we identify significant differences in regulatory activity occurring at a specific developmental stage between anterior and posterior *sc<sup>l+</sup>* populations (Figure 5-3).

**Figure 5-3 Distribution and overlap of non-promoter ATAC peaks in spatial comparisons of *sc<sup>l+</sup>* cellular contexts**



At the 10ss the anterior *sc<sup>l+</sup>* population is associated with only a third the number of genes in proximity to context specific non-promoter open chromatin peaks found in the posterior at the same stage. Additionally the majority of these genes (78%) are also found associated with posterior *sc<sup>l+</sup>* regions of open chromatin. This suggests

that within a single embryo at the 10ss the two *sc<sup>l</sup>*<sup>+</sup> populations experience significantly different levels of regulatory activity. High activity of the posterior program results in a greater number of genes in proximity to dynamically open regions of non-promoter chromatin, compared to the anterior. In the anterior this low number of genes associated with context specific regions of open chromatin, suggests that the regulatory program is not yet as active as in the posterior. This correlates with the earlier temporal comparison of these *sc<sup>l</sup>*<sup>+</sup> populations.

By the 20ss, the opposite regulatory activity is observed with the anterior *sc<sup>l</sup>*<sup>+</sup> population displaying a greater number of genes associated with dynamically open regions of non-promoter chromatin, compared to the posterior. This suggests that the regulatory activity of the anterior has now superseded that of the posterior either through a greater activity of a single regulatory program or through multiple programs occurring within this anterior *sc<sup>l</sup>*<sup>+</sup> population. The latter hypothesis correlates with the transcriptomes and imaging data for these populations, showing a range of behaviours in the anterior while the posterior remains strongly erythrogenic and vasculogenic.

### 5.3.3. Analysis of binding site motif enrichment within with context specific-ATAC peaks

We next analysed the ATAC-seq peaks specifically identified in different developmental contexts for statistically significant enrichment in DNA binding motifs. For instance ATAC-seq peaks, called in 10ss anterior ATAC-seq but not in 10ss posterior, represent a putative group of dynamically opened chromatin regions that most likely associate with specific *sc<sup>l</sup>*<sup>+</sup> anterior activity and not a *sc<sup>l</sup>*<sup>+</sup> posterior program.

For this analysis the JASPAR binding site analysis database was used<sup>365,366</sup>. Only binding sites for transcription factor orthologues that were identified by RNA-seq as expressed in each of these *scf*<sup>+</sup> populations were assayed. The *id* protein family is included in the JASPAR database as a transcription factor, despite this family lacking any DNA binding activity. For this analysis, despite expression of *id* family members in each of the *scf*<sup>+</sup> populations, we have removed this factor from our JASPAR mediated binding site search. We then searched for all the binding sites within each context-specific non-promoter open chromatin region dataset, as well as against total peaks as a background. The Fischer exact statistical test was performed to determine whether identified binding motifs were significantly enriched over background. By comparing the binding motifs, identified for each context, that have higher than expected presence, we identified putative upstream signatures that characterise each context (figure 5-4).

The majority of statistically significant enriched upstream transcription factors are shared by several or all contexts. This includes *scf* (*tal1*) itself, plus other key regulatory factor families such as *sox* and *fox* that are involved in a range of regulatory activities. Currently 20 *sox* genes have been identified in humans and mice and have been shown to play crucial roles in a range of developmental programs<sup>367,368</sup>. Sox17 has been demonstrated to be required for the maintenance of embryonic HSCs, possibly through action as a pioneer factor<sup>369-371</sup>.

In addition the binding site motifs for the haematopoietic and vascular *runx* and *kdrl* families were also identified as being enriched in the context specific non-promoter open chromatin regions for all four contexts. This suggests that despite these factors potentially bind to different open chromatin regions in these contexts, in all four contexts these key transcription factors are potentially play roles in directing the context specific active regulatory program.

A core of 5-7 transcription factor families (*usp*, *hif*, *tead*, *pdm1*, *gli*, *atf*, *fos*) was identified that show specific enrichment in early anterior and late posterior datasets. This is surprising as the analysis in section 5.2.1 and 5.2.2 suggest that these two populations are associated with dramatically different regulatory activity. The 10ss anterior shows the least regulatory accessible chromatin of the four contexts, while the 20ss posterior has been observed to be a reduction in open chromatin regions available for regulation following a highly accessible stage at 10ss. This core of families associated with the enriched binding sites of 10ss anterior and 20ss posterior specific non-promoter open chromatin peaks include factors previously demonstrated as contributing to the haematopoietic fate.

**Hif-** The hypoxia induced factor family are transcription factors that become activated in response to low oxygen conditions<sup>372</sup> and respond by driving a range of activities including angiogenesis<sup>373</sup>, embryonic haematopoietic development<sup>374</sup> and wound repair<sup>375</sup>.

**Tead-** Goode *et al.* identified TEAD factors in a similar approach as employed here, when investigating haematopoietic specification in embryonic stem cells (ESCs)<sup>376</sup>. TEAD factor activity is regulated by the hippo signalling pathway and has been shown to regulate haematopoiesis in *Drosophila* and ESCs<sup>376-378</sup>. CHIP-seq with TEAD4 in ESCs has shown that many TEAD4 binding sites overlap with SCL and LMO2 validated binding sites<sup>376</sup>.

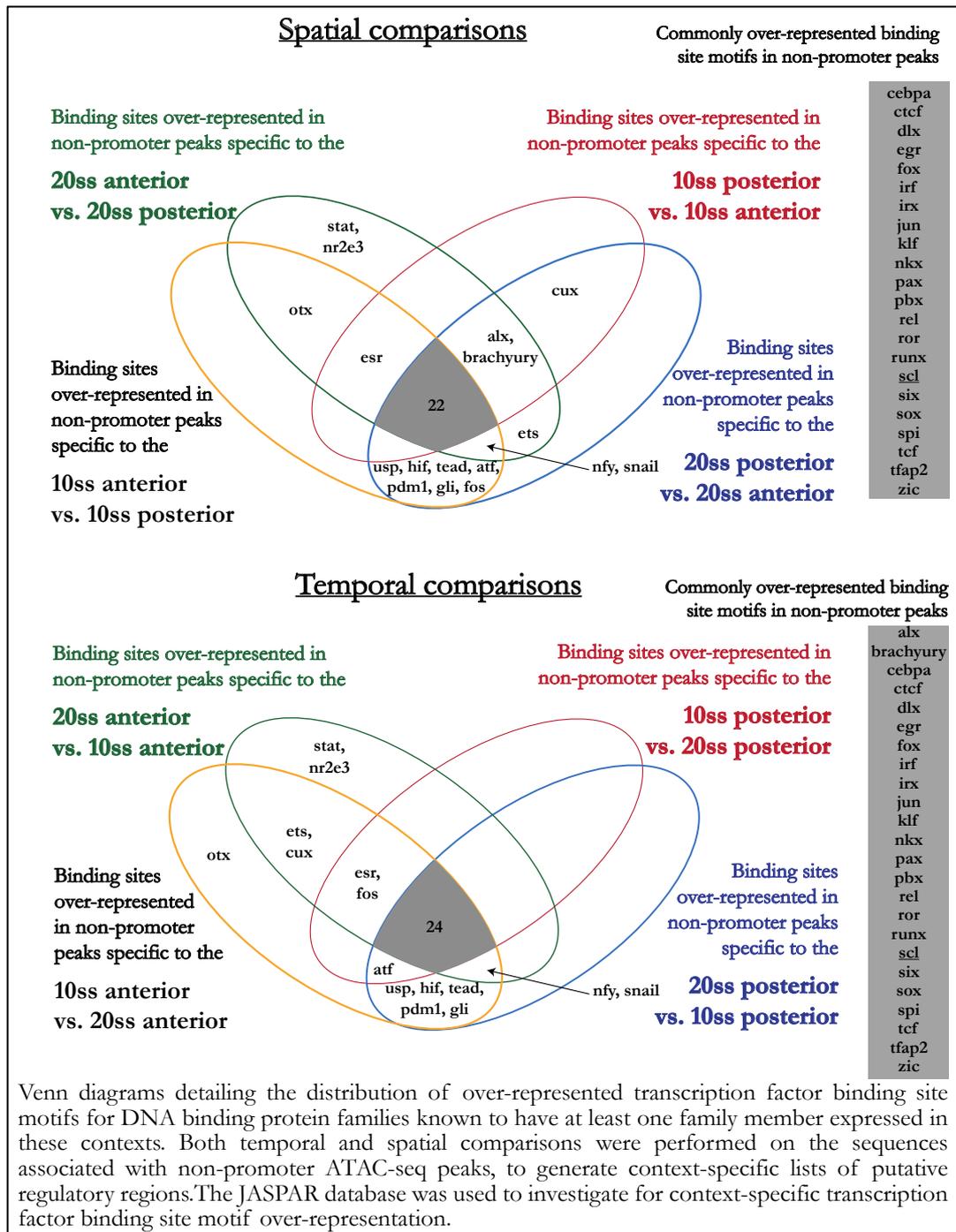
**Gli-** The Gli transcription factor family are involved in a range of developmental processes through their activity as effectors of the Hedgehog signalling pathway<sup>379</sup>. Knockdown studies have shown Gli1 to play a complex role in the regulation of haematopoietic stem cell number and myeloid development in mice<sup>380</sup>. Mouse knockouts of the *Ihh* ligand, which acts upstream of the Gfi proteins, causes 50% fatality at midgestation due to yolk sac abnormalities<sup>381,382</sup>.

Comparison to GO terms associated with enriched genes expressed by the 10ss anterior and the 20ss posterior *sc<sup>+</sup>* populations shows that both of these populations show enriched expression of genes annotated as being involved in myeloid development (Figure 4-16 and Figure 4-20). However key myeloid factors are also expressed in the posterior *sc<sup>+</sup>* population at the 10ss, yet this is not accompanied by significant enrichment of binding sites for the factors listed above in its' context specific non-promoter peaks.

One interpretation of this common binding site enrichment between these two *sc<sup>+</sup>* contexts could be that these factors contribute to the onset of a specific program only active in the early anterior of the embryo and the later in the posterior ICM. However this analysis is based on enrichment of binding sites for protein families- the activities of different family members can drive varied and even opposing cellular decisions. Thus at this point there is no evidence that these commonly enriched binding sites of the non-promoter peaks of the anterior and 20ss posterior are bound by the same members of these transcription factor families, or driving the same regulatory activity.

This binding site enrichment analysis also identifies context specific enrichment that is not shared with other *sc<sup>+</sup>* populations. Such enrichment may be the result of DNA-binding protein families such as *stat* and *nr2e3* being involved in activities specific to the 20ss posterior *sc<sup>+</sup>* population (see figure 5-4). However to draw such conclusions, the specific family member would have to be identified, confirmed to bind and exert regulatory activity at these putative regulatory sites, and contribute to context specific activity of that specific *sc<sup>+</sup>* population.

Figure 5-4 Venn diagram of context-specific binding site motif enrichment in non-promoter peaks



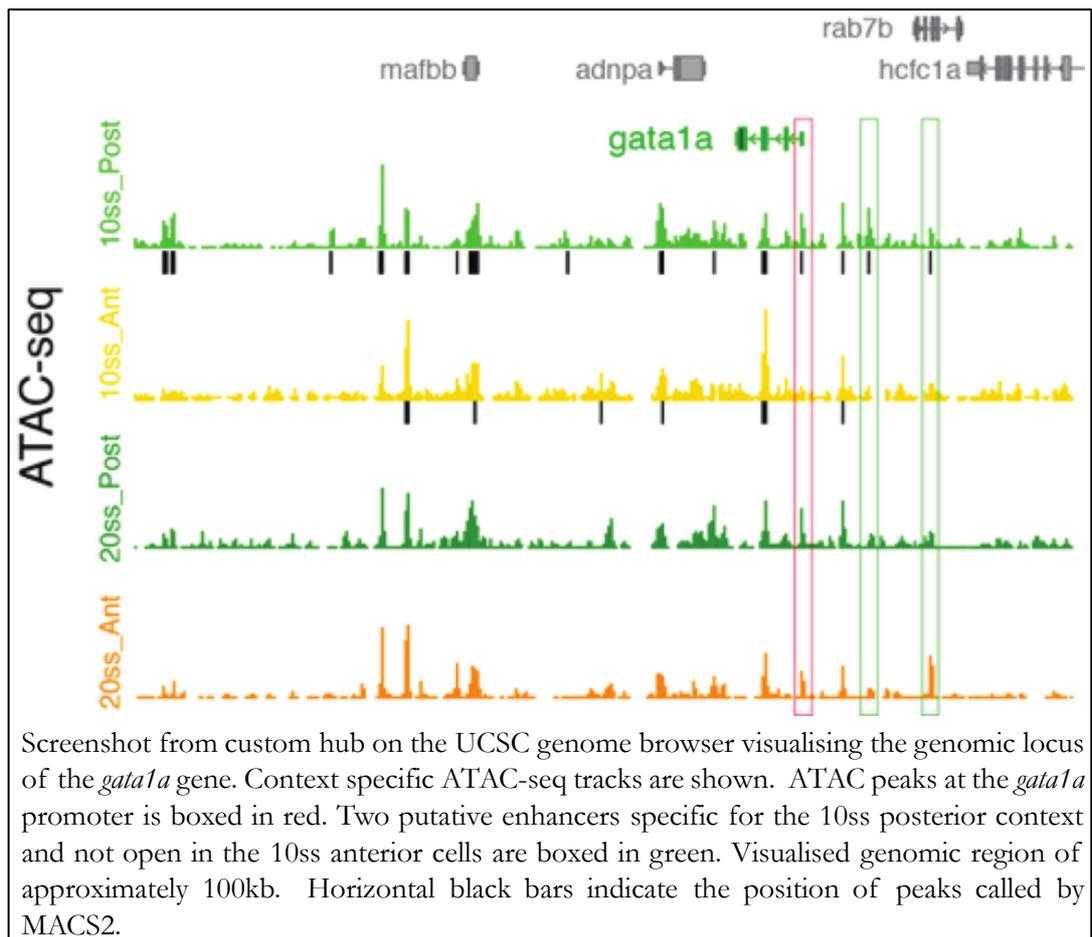
#### 5.4. Chapter 5 summary

To investigate the role of *scl*<sup>+</sup> early haematopoietic and vascular progenitors, open chromatin maps were produced for each of the cellular contexts studied by RNA-seq. I mapped reads from ATAC-seq to the zebrafish genome, called peaks and

associated these with putative genomic features based on proximity and expression level of proximal genes.

Figure 5-5 shows a genome browser image of ATAC peaks at the *gata1a* locus for each of the *scf*<sup>+</sup> contexts investigated. RNA-seq datasets confirm that *gata1a* is expressed in all four *scf*<sup>+</sup> populations but at significantly higher levels in the posterior especially at the 20ss. This figure demonstrates the power of combining these RNA-seq and ATAC-seq datasets to identify potential regulatory regions in proximity to genes of interest and relate these to context specific expression levels.

**Figure 5-5 Open chromatin maps at the *gata1* locus of each of the contexts studied.**



Comparison of ATAC-seq and RNA-seq datasets revealed that 90% of promoters that show open chromatin are associated with significant transcription in the 10ss posterior and both 20ss samples. In the 10ss anterior *scf*<sup>+</sup> cells only 40% of open promoters are related to transcription of the associated gene. This may suggest that

expression of these genes is primed and that the 10ss anterior *scf* population is more dynamic or developmentally “younger” than the other three studied *scf* contexts. The majority of promoters open in the 10ss posterior *scf*<sup>+</sup> population correlated with transcription of the associated gene, a pattern that is maintained to the 20ss stage. Together my *scf*<sup>+</sup> profiles suggest that the anterior population is initiating a transcriptional program at the 10ss, while the 10ss posterior *scf* population has already completed initiation of a regulatory program, which doesn’t dramatically vary by the 20ss.

Genome-wide maps of chromatin accessibility at non-promoter regions for each of these *scf*<sup>+</sup> populations were compared spatially and temporally. Non-promoter regions showing differential accessibility were anticipated to contain regions that contribute to the context specific cellular identity as putative regulatory regions.

Analysis of differences in open chromatin revealed that the majority of open chromatin sites outside of promoters in the 20ss posterior *scf*<sup>+</sup> cells are already open chromatin regions at the 10ss. This supports the transcriptomic data that suggests that the erythroid program is already engaged in the posterior *scf*<sup>+</sup> population, by the 10ss and then maintained to the 20ss. Comparison of ATAC-seq data for the anterior details a dramatic change in chromatin accessibility between 10 and 20ss. Only 10% of regions of open chromatin found in the 20ss anterior *scf*<sup>+</sup> population are also found to be open in the 10ss anterior population, which suggests a dramatic change in the regulatory landscape occurs during this time.

Initial binding site analysis has been performed, identifying transcription factor binding sites over-represented in non-promoter open chromatin regions for each *scf*<sup>+</sup> context studied. Combination of binding site analysis with transcriptomic data offers the possibility of proposing putative regulatory networks for each of these early

haemangiogenic contexts, but requires further investigation before conclusions can be drawn.

Combination of transcriptomic and chromatin accessibility datasets for each of these populations provides an excellent tool for identifying putative enhancers with context specific activity. Further analysis of these datasets could greatly contribute to the construction of a dynamic gene regulatory network model, for primitive haematopoiesis in vertebrates.

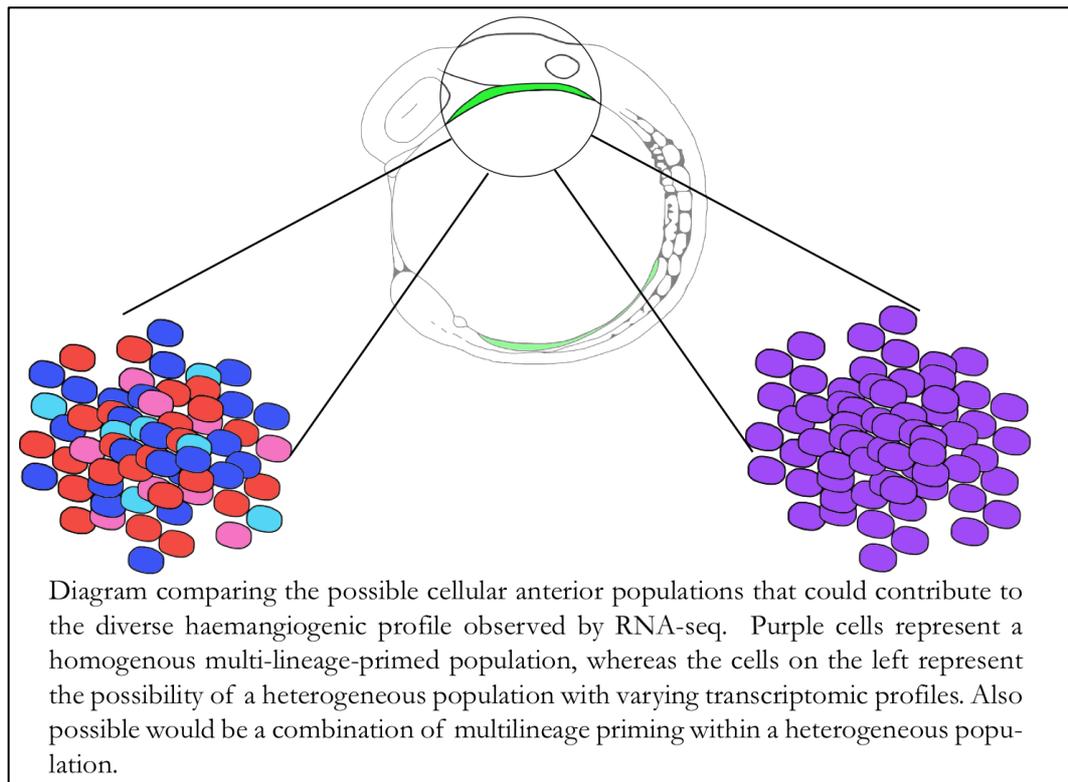
## 6. Investigating early anterior diversity of *scf* expressing cells

### 6.1. Introduction

RNA-Seq and ATAC-Seq profiles for the 10ss anterior *scf* expressing cells described the expression of key regulator genes for multiple lineages. By the 20ss both anterior and posterior *scf* populations have begun to develop into multiple haematopoietic or vascular lineages. This tagged *scf* transgenic reporter line can be used to follow the dynamics of *scf*<sup>+</sup> cells as shown in figures 3-6, 3-7 and 3-8. The behaviour of different embryonic populations can be evaluated by crossing the *scf* transgenics to another line that fluorescently marks a different population of cells in the developing embryo. Such comparisons provide information on the interactions and precise location of the *scf*<sup>+</sup> population *in vivo*. Using cutting edge imaging techniques and processing pipelines, this spatial and temporal data is available at a previously unavailable resolution.

Initially *scf*<sup>+</sup> sub-populations, at the 20ss, were described in relation to other key cell populations using *in vivo* imaging techniques to visualise co-expressing and distinct clusters of cells. The diversity in the 10ss anterior profiles could result from multi-lineage-priming or a mix of *scf*<sup>+</sup> sub-populations (see figure 6-1). Multi-lineage-priming is the expression of genes required for a variety of lineages, before the progenitor has differentiated or become committed to a lineage. This process has been described previously in cell cultures of haematopoietic precursors<sup>383,384</sup> and would suggest that the 10ss anterior *scf*<sup>+</sup> population is a homogeneous collection of cells. To discern between multi-lineage priming and the divergence of an early blood and vascular progenitor population, this population was spatially compared to LPM and endothelial populations within the 10ss embryo.

Figure 6-1 Schematic of potential cellular diversity that could produce varied biological GO profiles.



The spatial description of  $scf^+$  sub-populations by microscopy techniques was combined with single cell sequencing of  $scf^+$  cells isolated from the 10ss anterior. Single cell sequencing provided the transcriptome of individual cells, allowing for variation in expression patterns to be investigated and confirmation of co-expression of factors within a single cell. Transcriptional data at cellular resolution can be used in the future to determine gene regulatory networks of these developmentally important populations.

## 6.2. Visualising $scf$ sub-populations *in vivo*.

An initial investigation into the early heterogeneity of the anterior  $scf^+$  population was carried out using high resolution imaging of embryos produced from crosses of three transgenic lines. Each of these transgenic lines show tissue-specific expression of a reporter gene expressed from a transgenic BAC construct that has been stably integrated into the zebrafish genome. This is a highly useful approach for observing

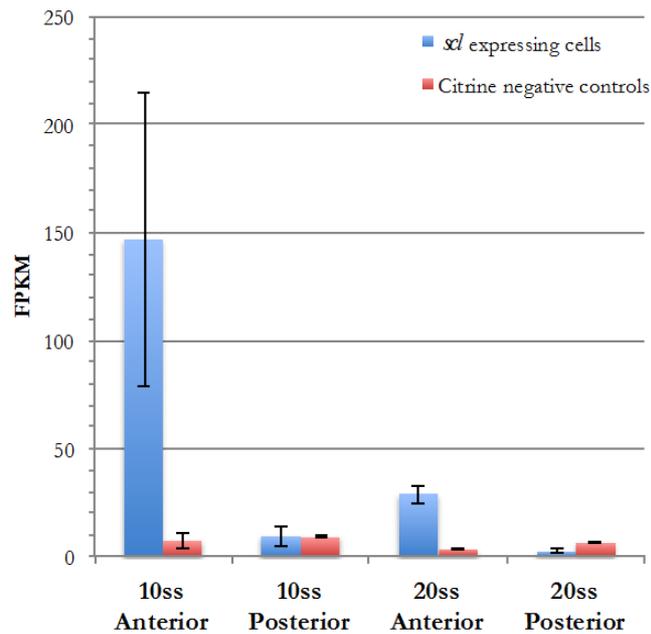
overlapping and distinct cell populations but is reliant on the production and detection of fluorescent reporter proteins. These transgenic reporter transcripts and proteins will not be subject to the same regulatory activities that the endogenous gene products will experience and protein half-life can be significantly greater than the half life of its transcript. These issues can cause false positives to be observed as cells cease to transcribe from the loci but the fluorescent reporter can persist. False negatives can also arise as a cell begins to express a loci but the reporter protein requires time to correctly fold and produce detectable levels of fluorescence.

### 6.2.1. *scf* populations related to the lateral plate mesoderm.

GATA5 is a transcription factor that is part of the GATA family that bind to the DNA sequence “A/TGATAA/T” using two zinc finger domains. Expression of *gata5* is first observed in endodermal progenitors at dome stage and moves into the anterior lateral plate mesoderm (ALPM)<sup>385</sup>. The *scf*<sup>+</sup> population derives from this *gata5*<sup>+</sup> lateral plate mesoderm in the anterior of the embryo. Reiter *et al.* demonstrated that *gata5* is involved in regulating the development of the endoderm and specifically the vertebral heart<sup>287</sup>. Other studies have also shown that *gata5* plays a role in the development of the zebrafish gut tube, liver, pancreas, thyroid and thymus- all of endodermal origin<sup>386</sup>.

In this investigation I used a *gata5* reporter line available at a collaborator’s institution to identify the lateral plate mesoderm. The Tg(*gata5*BAC:*L.A.-GFP*) zebrafish line has been generated by BAC transgenesis in the lab of S. Fraser, University of South California, by Dr Le Trinh. Expression of the *L.A.-GFP* reporter gene is driven by the *gata5* locus contained within this BAC.

Figure 6-2 Expression levels of *gata5* in early *scf*<sup>+</sup> populations.



Histogram showing the level of *gata5* expression within the *scf* expressing cells and negative controls of early zebrafish development. Variation within replicates as standard error is represented by error bars.

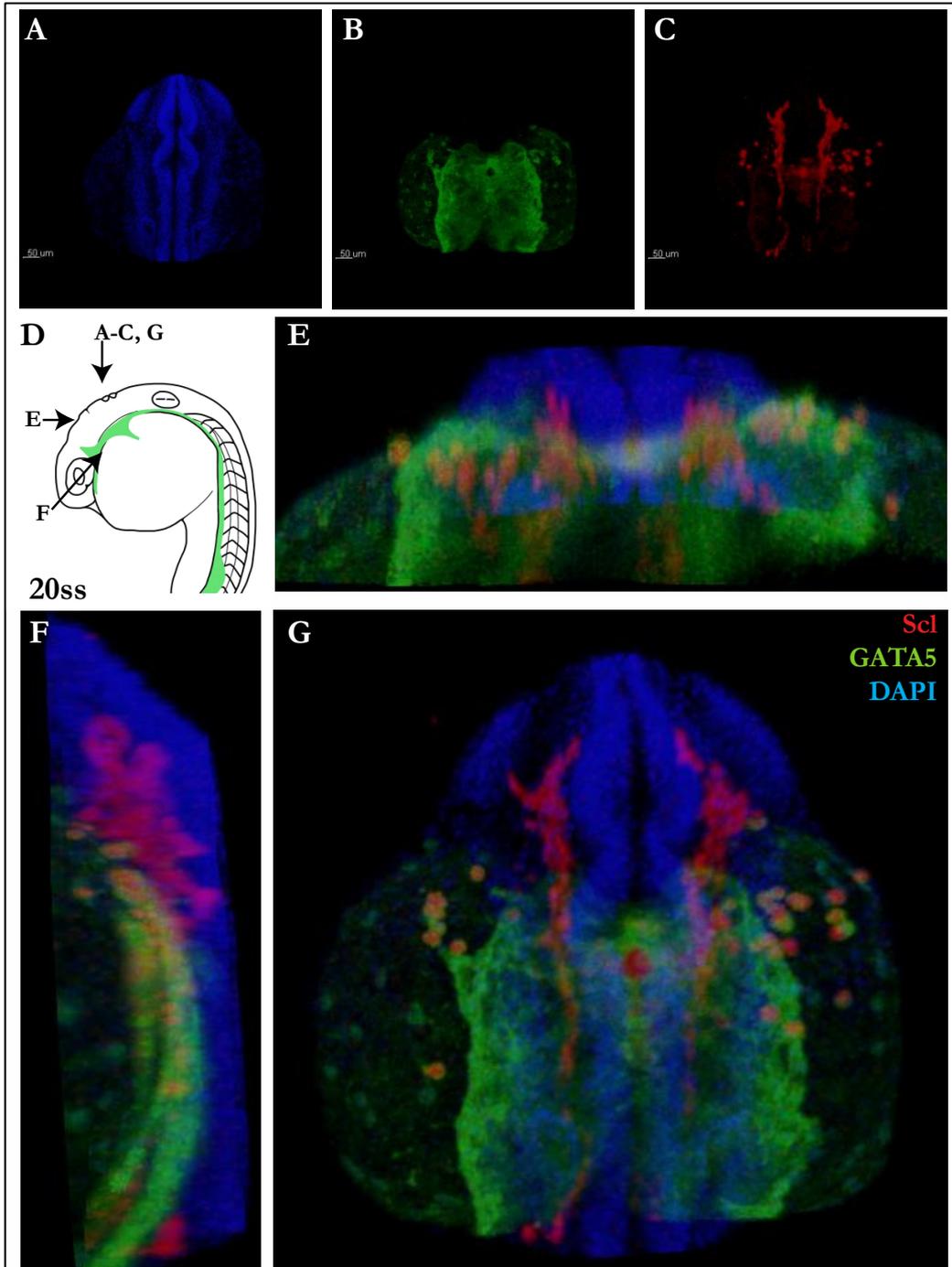
L.A.-GFP is a fusion protein of the N-terminal motif of the Abp120 protein from *Dictyostelium discoideum* to the C-terminal of a GFP reporter protein. The 17 amino acids from the start of Abp120 have been shown to be sufficient to bind actin filaments. Upon fusion to GFP, the fluorescent reporter is localised to actin filaments<sup>387</sup>. This is advantageous for the *gata5* reporter line as *gata5*<sup>+</sup> cells undergo significant changes in morphology as they differentiate from the mesoderm. In this investigation L.A.-GFP fluorescence enables *gata5*<sup>+</sup> cells of the mesoderm to be distinguished from *gata5*<sup>+</sup> cells moving over the yolk after emerging from the mesoderm, through visible differences in morphology. The RNA-seq data I have produced correlates with an anterior overlap of these two populations (see figure 6-2). *gata5* transcripts show 20 and 8.5 fold enrichment over context-matched *scf* cells in the 10ss and 20ss anterior samples respectively. This suggests a high level of co-expression with *scf* specifically in the anterior. *gata5* expression in the *scf* population is

not insignificant suggesting that there is not complete overlap between the two populations.

The early comparison between the lateral plate marker and *scl*<sup>+</sup> cells was used to annotate the location of the *scl*<sup>+</sup> populations previously investigated in this study, in relation to the LPM. Double positive embryos produced from a cross between the Tg(*gata5*BAC:*la*-GFP) line and the Tg(*scl*BAC:*scl*-flag-*tev*-*avi*-2A-*citrine*-*sv40pA*) transgenic line were studied at the 10ss and the 20ss.

By 20ss the *scl* expressing cells are seen as 3 distinct populations- individual cells over the yolk, cluster of cells at the midline and a pair of bilateral stripes either side of the neural tube (figure 6-3). *gata5:la-gfp*<sup>+</sup> marks the ALM and is co-expressed with *scl* in the cells over the yolk and posterior cells of the anterior bilateral stripes. L.A.-GFP protein is excluded from the central cluster of *scl* expressing cells within the developing heart field, though is highly expressed in the proximal cells. The anterior-most cells of the anterior bilateral stripes of *citrine* expression do not express *gata5:la-gfp*<sup>+</sup>. This population diverges into strings of *citrine*<sup>+</sup> cells enveloping the developing eye- this *scl* subpopulation may be the origins of the head vasculature. *Citrine*<sup>+</sup> cells from this position have been observed to develop into vessels (see figure 3-8). These observed expression patterns correlate with the known functions of *gata5*. GATA5 is required for myeloid cell differentiation<sup>286</sup> and correlates with the co-expression with *scl* in individual cells over the yolk. GATA5 has also been demonstrated to be required for migration of cells into the developing heart and for differentiation of cardiomyocytes<sup>285,286,388</sup>. *gata5:la-gfp*<sup>+</sup> cells surrounding the *citrine*<sup>+</sup> *gata5:la-gfp*<sup>-</sup> cells of the developing endocardium, are positioned where cardiomyocytes would be predicted to lie<sup>388</sup>.

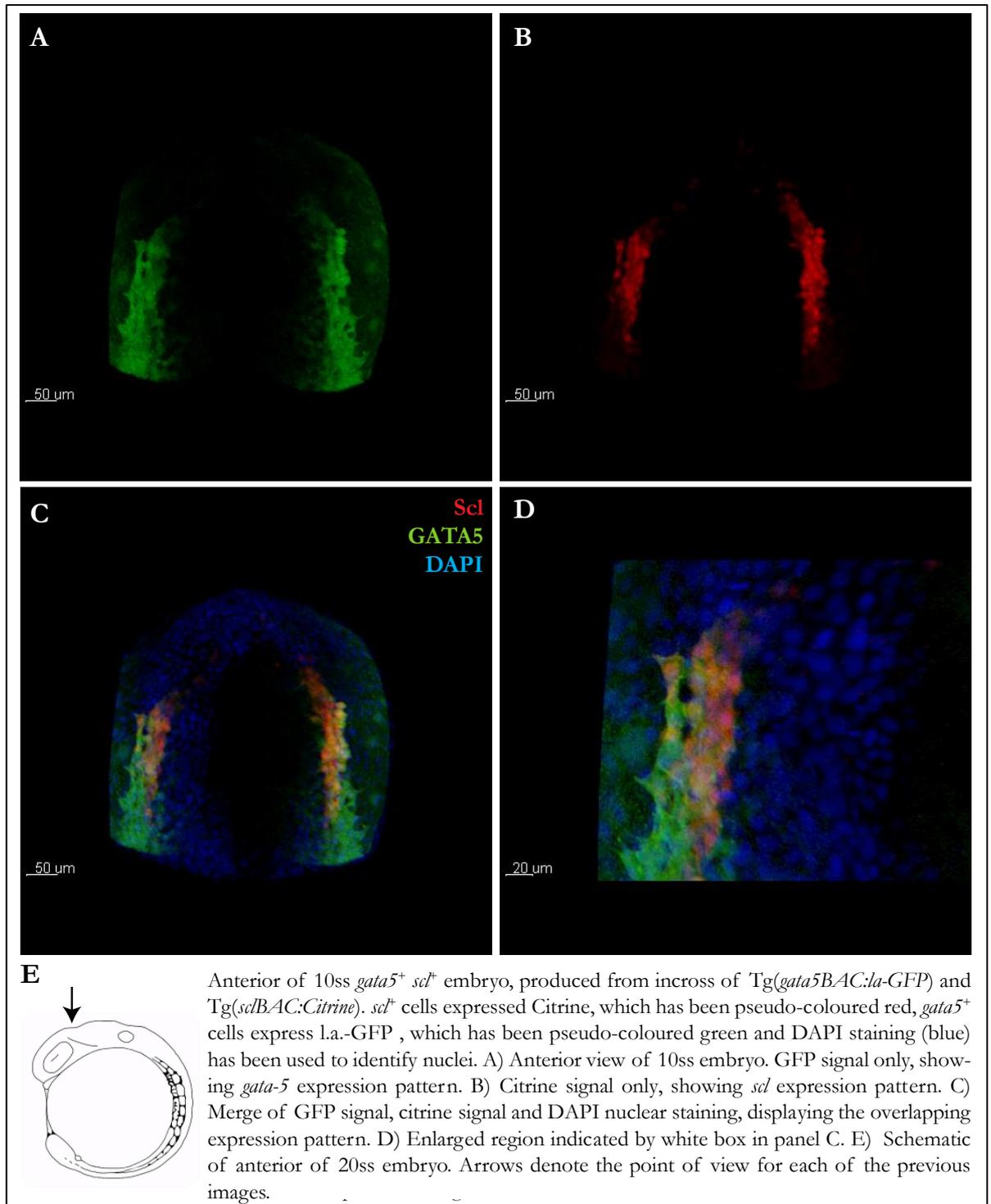
Figure 6-3 Hyperspectral imaging of the 20ss embryo with *gata5*<sup>+</sup> and *scf*<sup>+</sup> populations labelled.



Anterior of 20ss *gata5*<sup>+</sup> *scf*<sup>+</sup> embryo, produced from incross of Tg(*gata5*BAC:*la-GFP*) and Tg(*scf*BAC:*Citrine*). *scf*<sup>+</sup> cells express Citrine, which has been pseudocoloured red, *gata5*<sup>+</sup> cells express *la-GFP*, which has been pseudo coloured green and DAPI staining (blue) has been used to identify nuclei. A) DAPI nuclear signal only, showing the dorsal view of anterior of the 20ss embryo. B) L.A.-GFP signal only, showing *gata-5* expression pattern C) Citrine signal only, showing *scf* expression pattern. D) Schematic of anterior of 20ss embryo. Arrows denote the point of view for each of the following images. E-G) Merge of GFP signal, Citrine signal and DAPI nuclear staining, displaying the overlapping expression patterns of *gata-5* and *scf*. E) Transverse reconstruction down the neural tube, dorsal to top. F) Lateral reconstruction, anterior to top, dorsal to right. G) Merge of GFP signal, citrine signal and DAPI nuclear staining, displaying the overlapping expression pattern

*citrine*<sup>+</sup> cells have been observed to migrate from the bilateral stripes of *citrine*<sup>+</sup> (*scl*) expression into the central heart field (see figure 3-8). These *scl*<sup>+</sup> cells would be passing through a *gata5*<sup>+</sup> population, permitting the opportunity for intercellular signalling, that could guide *scl*<sup>+</sup> cell migration. *gata5* has not been reported to contribute to endothelial development, which correlates with the lack of its expression in the anterior *scl*<sup>+</sup> cells that potentially contribute to cranial vasculature. It was investigated whether these *scl*<sup>+</sup> subpopulations, with differential *gata5* expression, are already formed at the 10ss. To address this question these double transgenics were studied at an earlier developmental stage (figure 6-4). At the 10ss *gata5:la-gfp*<sup>+</sup> cells exist as a pair of bi-lateral stripes indicating that the ALM has yet to return to the midline, and thus at this stage *gata5* acts as a marker for the LPM. *citrine* expression closely overlaps in the anterior, indicating that at this stage the *scl*<sup>+</sup> population has yet to migrate out of the LPM. *citrine*<sup>+</sup> cells are also *la-gfp*<sup>+</sup> in my 10ss investigations, though *citrine*<sup>+</sup> *gata5:la-gfp*<sup>+</sup> cells are observed to express lower levels of *la-gfp* than *citrine*<sup>-</sup> *gata5:la-gfp*<sup>+</sup> cells. This suggests that the presence of GATA5 is not responsible for diversification of the *scl*<sup>+</sup> population at this stage but that a gradient of *gata5* expression may contribute to cell fate decisions of the early *scl*<sup>+</sup> population.

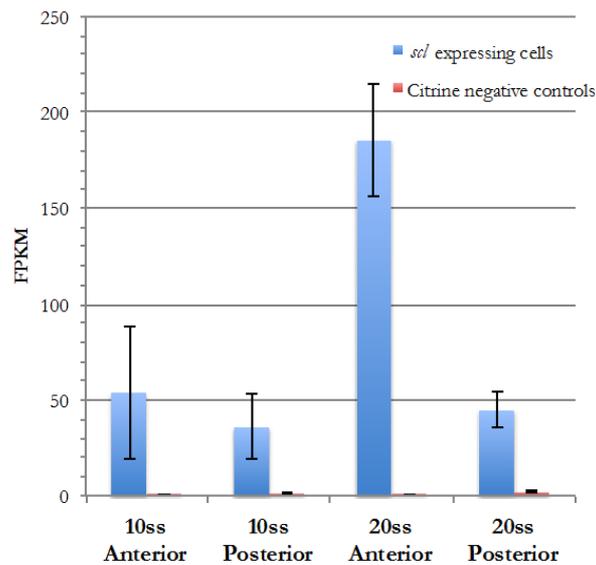
Figure 6-4 Hyperspectral imaging of a double transgenic 10ss embryo produced from a cross of the *Tg(gata5BAC:la-GFP)* and the *Tg(sclBAC:scl-flag-tev-avi-2A-citrine-sv40pA)* transgenic lines.



### 6.2.2. *scl*<sup>+</sup> populations in relation to endothelial progenitor cells.

*Kdr1* is commonly used as a molecular marker of endothelial cells in zebrafish and is expressed throughout the vascular system<sup>389</sup>. This VEGF receptor is required for correct formation of key blood vessels, with loss and delayed development of the dorsal aorta and intersegmental vessels in *kdr1* mutants<sup>390</sup>.

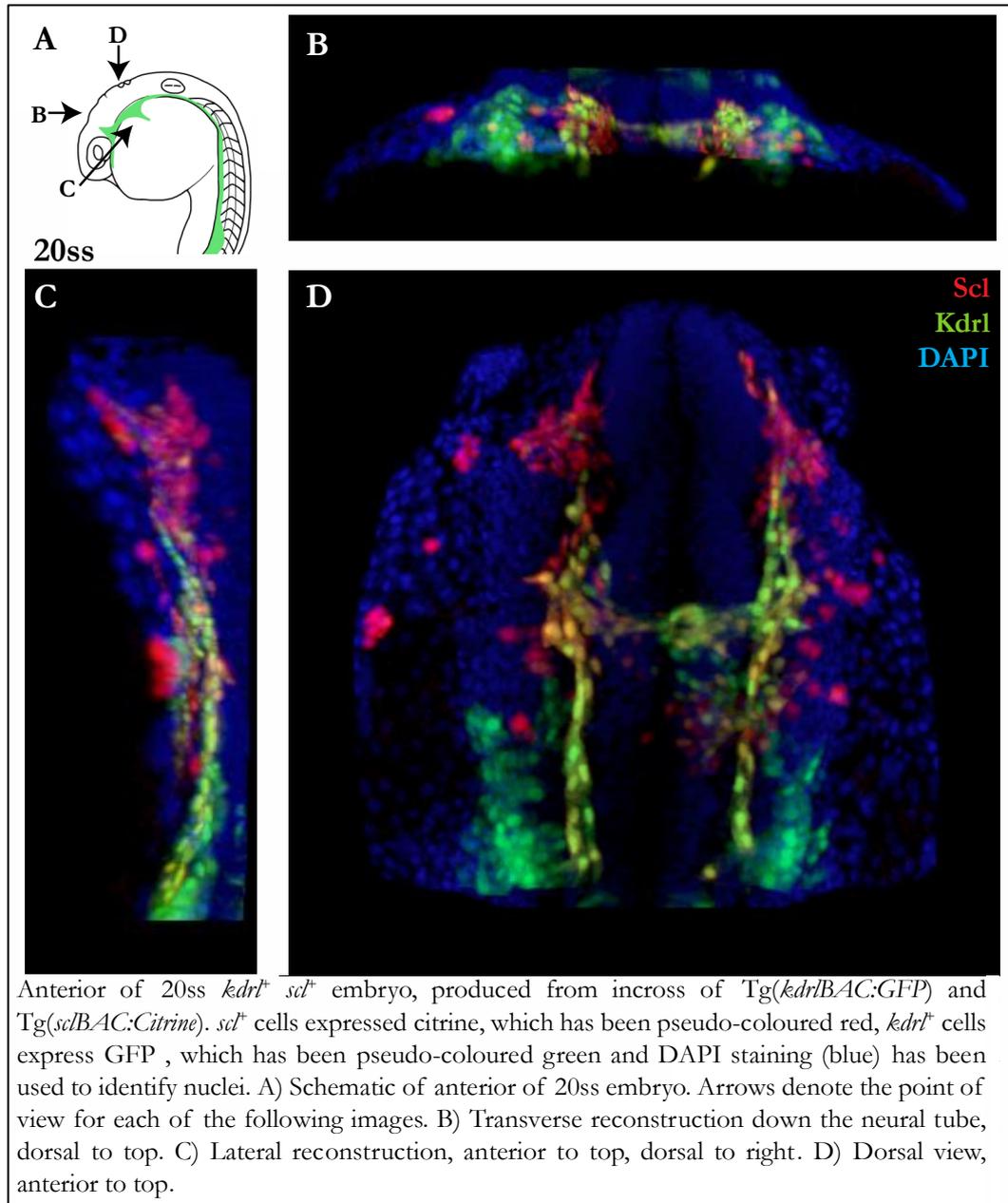
**Figure 6-5 Expression of *kdr1* in early *scl* expressing populations**



Histogram showing the level of *kdr1* expression within the *scl* expressing cells and negative controls of early zebrafish development. Variation within replicates as standard error is represented by error bars.

I used a *Tg(kdr1BAC:gfp)* reporter line available at a collaborator's institution to introduce an endothelial marker alongside the *scl* signal in the *Tg(sclBAC:scl-flag-*tev-avi-2A-citrine-sv40pA*)* transgenic line. This *Tg(kdr1BAC:gfp)* reporter line had previously been characterised and determined to accurately mimic the expression of endogenous *kdr1*, by Dr Le. Trinh, in the Fraser lab at the University of South California. The combination of these two fluorescently marked populations would enable *scl*<sup>+</sup> endothelial progenitors to be visualised and potentially mark a subpopulation of *scl*<sup>+</sup> cells in the anterior of the embryo.

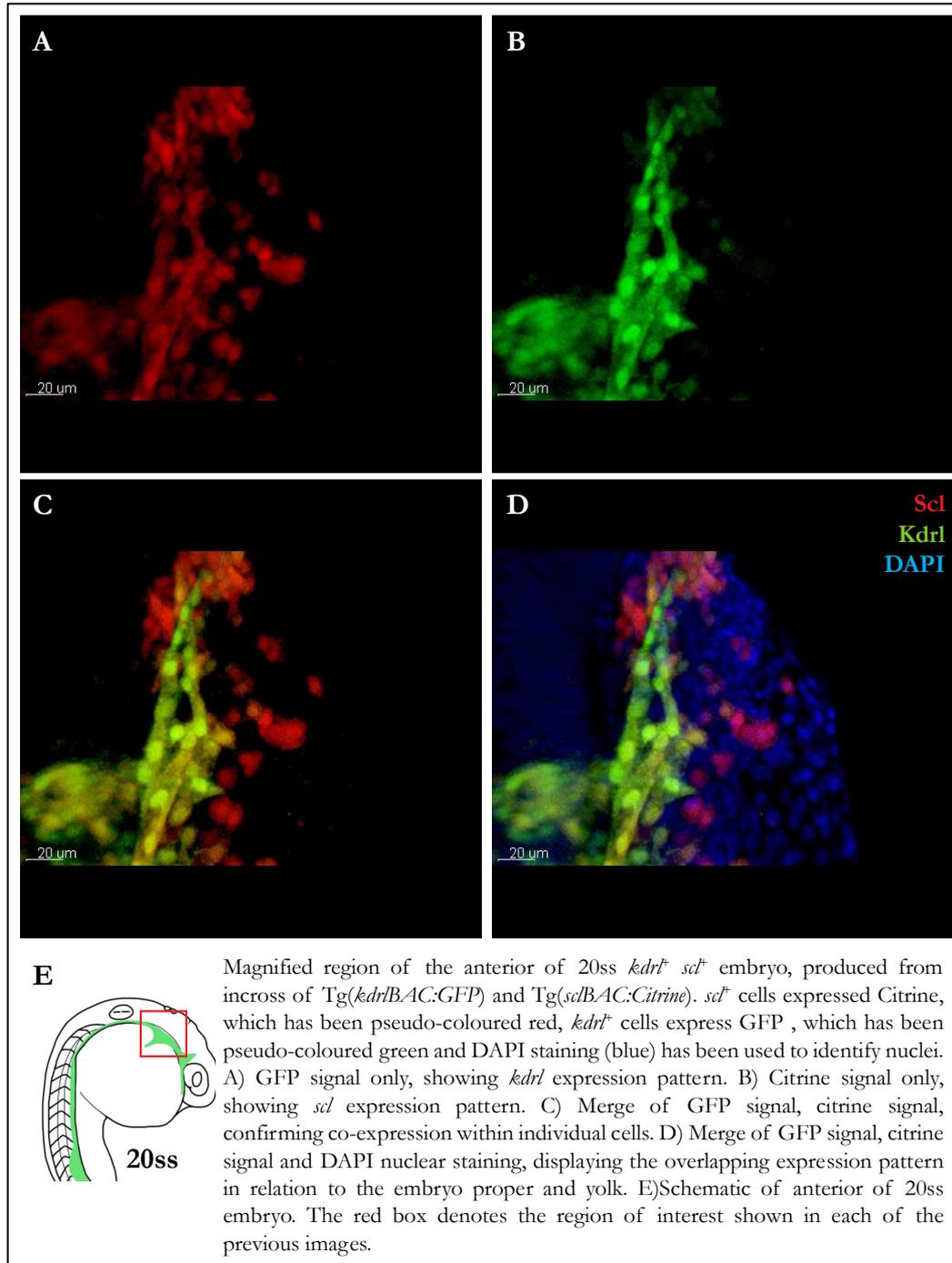
Figure 6-6 Double transgenic embryo produced from a cross of *Tg(kdr1BAC:gfp)* and *Tg(sclBAC:scl-flag-tev-avi-2A-citrine-sv40pA)* transgenic lines, fluorescently marking the *kdr1* and *scl* expressing populations at the 20ss.



From the RNA-seq data of the early *scl* expressing populations, *kdr1* expression is highly enriched in *scl<sup>+</sup>* cells of each of the contexts studied (figure 6-5). Expression is high in all four *scl<sup>+</sup>* populations and at it's highest in the 20ss anterior. This suggested a high overlap of the *kdr1<sup>+</sup>* and *scl<sup>+</sup>* populations during the early stages of blood and vascular development.

Embryos produced from in-crosses of  $Tg(kdrlBAC:GFP)$  and  $Tg(sclBAC:scl-FLAG-TeV-Avi-2A-citrine)$ , were fluorescently sorted and double positive embryos imaged at the 10 and 20ss.

**Figure 6-7 Magnified image of double transgenic embryo, illustrating the variation in *kdrl* and *scl* expression patterns at the 20ss.**



In the 20ss anterior the *scl*<sup>+</sup> and *kdrl*<sup>+</sup> populations significantly overlap, however distinct populations of both are identifiable (figures 6-6 and 6-7). GFP fluorescence

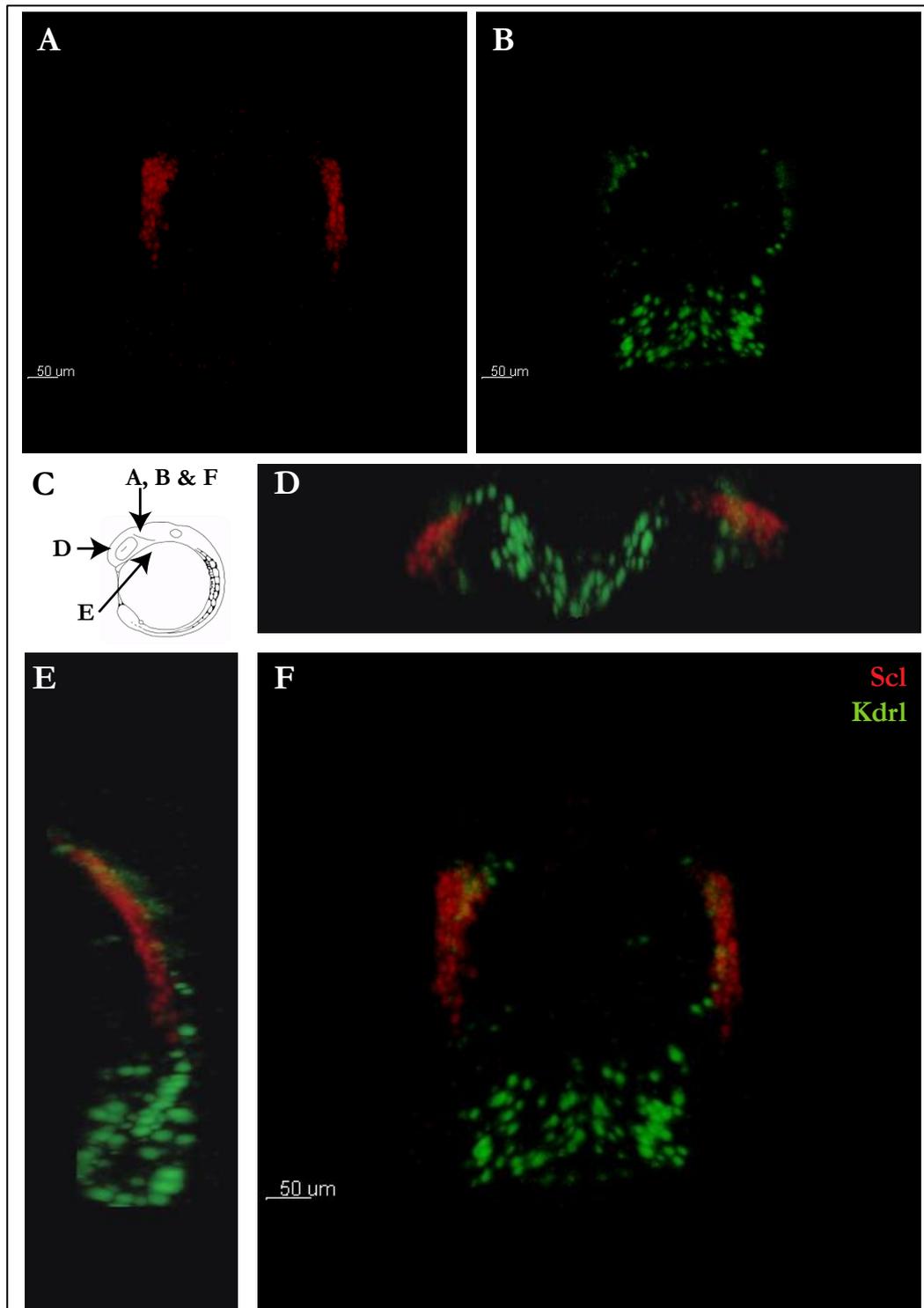
driven by the *kdrl* regulatory regions displays a range of expression levels, unlike *citrine* expression, driven by *scl* regulatory regions, which is observed to be expressed at a consistent level across the 20ss anterior population. *kdrl* and *scl* are co-expressed in the bilateral stripes that extend along the anterior: posterior axis of the anterior of the embryo. Expression of *kdrl* is greatest at the posterior of the anterior, where *scl* expression is minimal. *kdrl*<sup>+</sup> *scl*<sup>-</sup> cells were observed in a pair of populations that lie further laterally to the bilateral double positive populations.

Cells expressing *scl* but not *kdrl* are visible over the yolk. These have been proposed to be early myeloid progenitors, thus *kdrl*-controlled processes are not required in these cells. *kdrl* and *scl* are co-expressed in the cells of the developing endocardium concurrent with evidence that *kdrl* is required for valve formation within the heart<sup>391</sup>.

Cells at the anterior that are in proximity to the eye also co-express *scl* and *kdrl*, though *kdrl* is at significantly lower levels than in the endocardium. This is surprising as this *scl*<sup>+</sup> populations has been observed to contribute to cranial vasculature, thus expression of a key endothelial gene such as *kdrl* would be expected. Chung *et al.* demonstrate in primary culture models of definitive haemangiogenesis, that *scl* expression is induced within *kdrl*<sup>+</sup> cells and that endothelial cells developed from *kdrl*<sup>+</sup> progenitors<sup>19</sup>. The presence of this vasculogenic *scl*<sup>+</sup> *kdrl*<sup>low</sup> population suggests that during primitive haematopoiesis the regulatory network may differ in order from that of definitive haematopoiesis.

To investigate whether these different expression patterns between *kdrl* and *scl* were present earlier in haemangiogenic development, double positive embryos from in-crosses of Tg(*kdrl*BAC:*gfp*) and Tg(*scl*BAC:*scl-flag-tev-avi-2A-citrine*) were imaged at the 10ss (figure 6-8). *scl*<sup>+</sup> cells are observed as a pair of populations either side of the anterior: posterior midline, posterior to the developing eye.

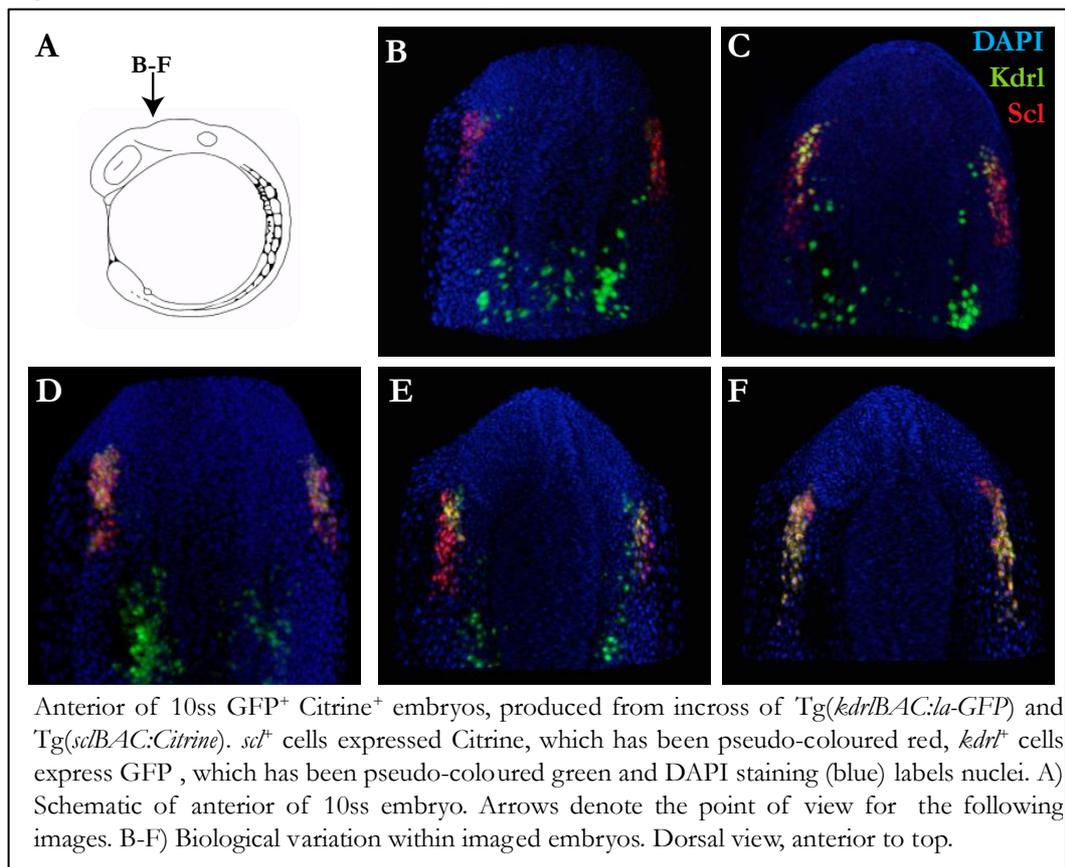
Figure 6-8 Expression of fluorescent reporters for *scf* and *kdr1* populations at the 10ss.



Anterior of 10ss GFP<sup>+</sup> Citrine<sup>+</sup> embryos, produced from incross of Tg(*kdr1*BAC:GFP) and Tg(*scf*BAC:Citrine). *scf*<sup>+</sup> cells expressed Citrine, which has been pseudo-coloured red, *kdr1*<sup>+</sup> cells express GFP, which has been pseudocoloured green. A) Citrine signal only, showing *scf* expression pattern. B) GFP signal only, showing *kdr1* expression pattern. C) Schematic of anterior of 10ss embryo. Arrows denote the point of view for the other panels. D-F) Merge of GFP signal and Citrine signal displaying the overlapping and distinct expression patterns. D) Transverse reconstruction down the neural tube, dorsal to top. E) Lateral reconstruction, anterior to top, dorsal to right. F) Dorsal view, anterior to top.

*kdr1*<sup>+</sup> cells form two pairs of bilateral populations; the most anterior of which overlaps to varying extents with the *scl*<sup>+</sup> population, and a more medial pair of *kdr1*<sup>+</sup> populations that do not express *scl*. The number of *kdr1* expressing cells is small at the 10ss and cells are dispersed, allowing individual cells to be easily resolved. In the most anterior labelled population a mixture of *scl*<sup>+</sup> *kdr1*<sup>+</sup>, *scl*<sup>+</sup> *kdr1*<sup>-</sup> and *scl*<sup>-</sup> *kdr1*<sup>+</sup> cells exist. This mixed population may represent a transformation of one developmental program to another or the emergence of two distinct populations from a common progenitor.

**Figure 6-9 Variation of *kdr1* expression patterns in the 10ss anterior.**



Unlike the 10ss expression pattern for *gata5* and *scl*, significant biological variation was identified in the spatial arrangement of early *kdr1*<sup>+</sup> cells as shown in figure 6-9. This variation was deemed not to be a result of largely differing developmental stage as embryos were checked for somite number and produced from a single lay. Due

to the significant differences in overlap of expression patterns it is unclear the degree to which *kdrl* expression could contribute to cellular variation within the *scf*<sup>+</sup> population.

### 6.3. Transcriptional profiles of individual anterior *scf*<sup>+</sup> cells at the 10ss.

#### 6.3.1. Overview of single cell sequencing results

10ss Citrine<sup>+</sup> cells were isolated as previously described and submitted to single cell RNA extraction and subsequent sequencing. 94 individual *scf*<sup>+</sup> cells were processed with the aim of describing at a cellular level the transcriptome of this early *scf*<sup>+</sup> population. Through this cellular approach, variation between cells can be resolved and described.

The majority of the cells were successfully processed and yielded between 0.5 -2 million mapped reads per sample, see Figure 6-10.

**Figure 6-10 Single cell RNA sequencing read data per sample.**

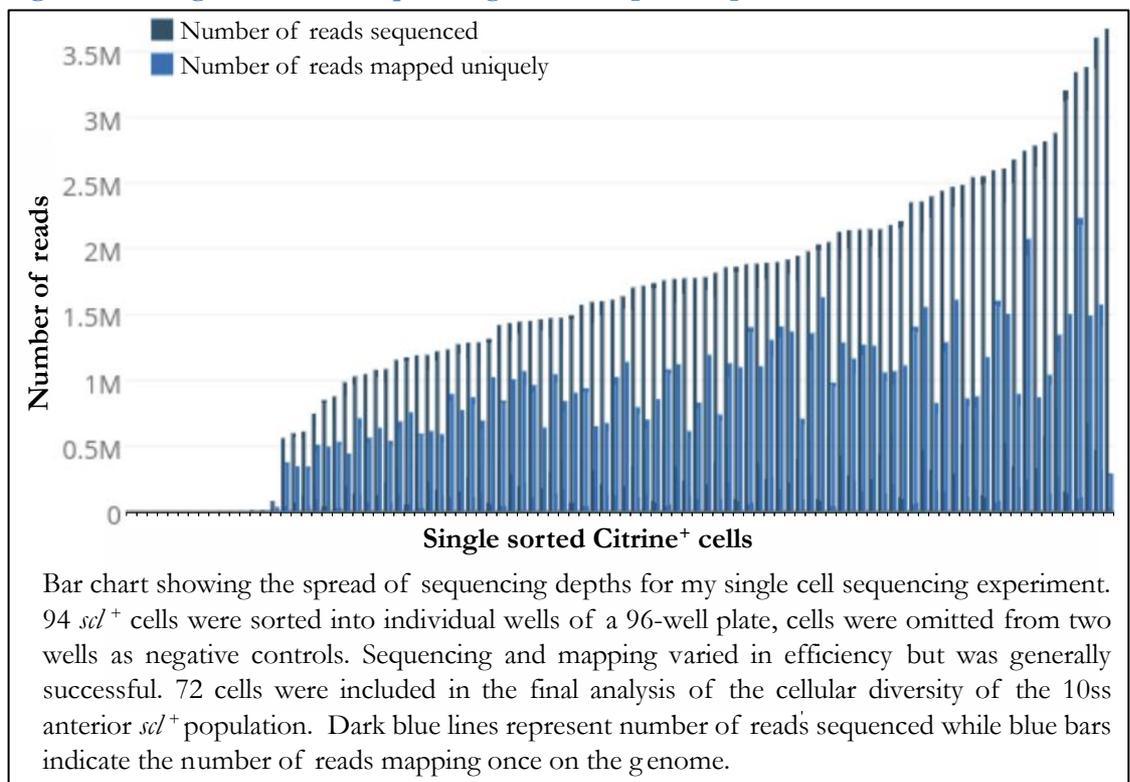
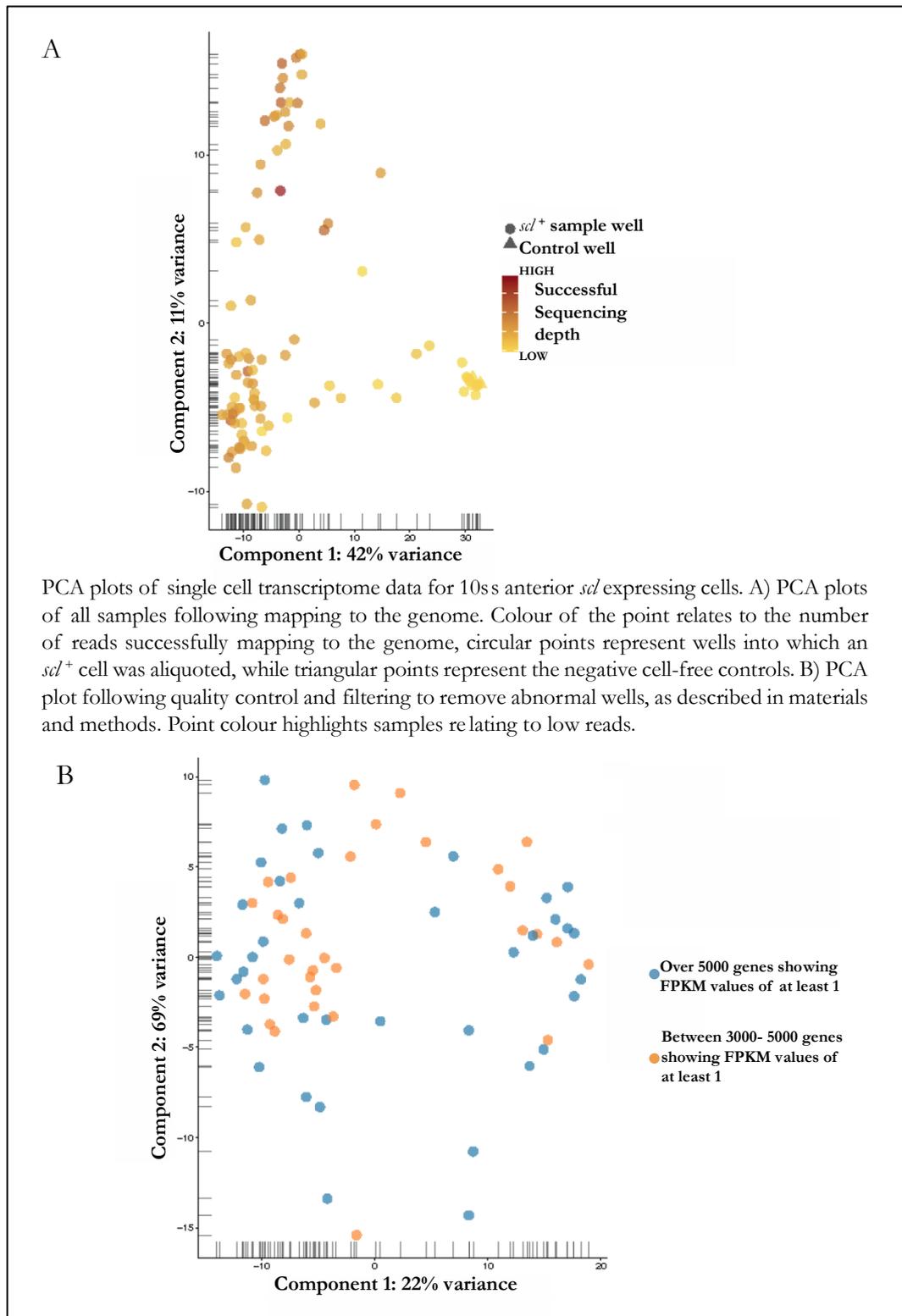


Figure 6-11 PCA analysis of single cell transcriptome data, before and after quality control and filtering.



Initial PCA analysis of the sequencing data suggested the presence of three cellular profiles as detailed in plot A, Figure 6-11. The cluster of sample on the right of plot A, Figure 6-11, correspond to wells were RNA extraction failed, sequencing depth

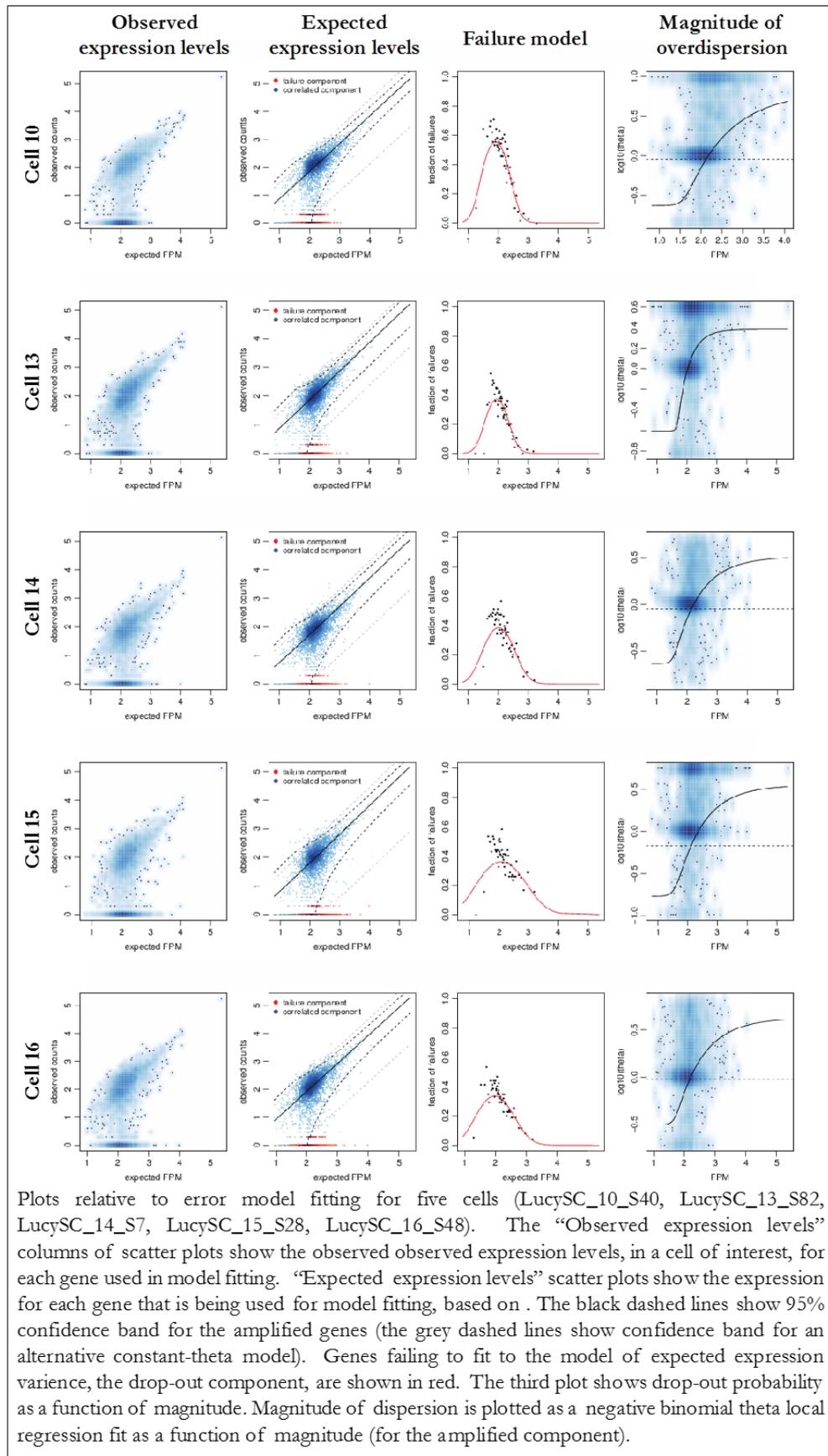
was minimal or the cell was a control cell, in which no cell had been aliquoted. Quality control was carried out and samples were omitted based on lack of sequencing depth, lack of ERCC DNA normalisation and poor genome coverage following mapping. PCA analysis was carried out after this refinement of the dataset and fails to show any discernable clusters of cells. It is possible that there may be two clusters but a greater number of samples would be required for this PCA analysis to resolve sub-populations.

### 6.3.2. Early anterior subpopulation transcriptomes.

In order to investigate whether subpopulations or groups of co-expressing genes could be identified, the R package PAGODA (v.1.99) was used to evaluate the variability of gene expression between single cell transcriptomes, in relation to annotated pathways and in *de novo* detected gene sets<sup>392,393</sup>. Single cell error models were fitted by determining effective sequencing depth, dropout rate (Poisson), and amplification noise for each cell. These are estimated as a mixture of a negative binomial (NB) (signal) and Poisson (drop-out) as described in Fan *et al.*, 2016<sup>393</sup> (see figure 6-12)

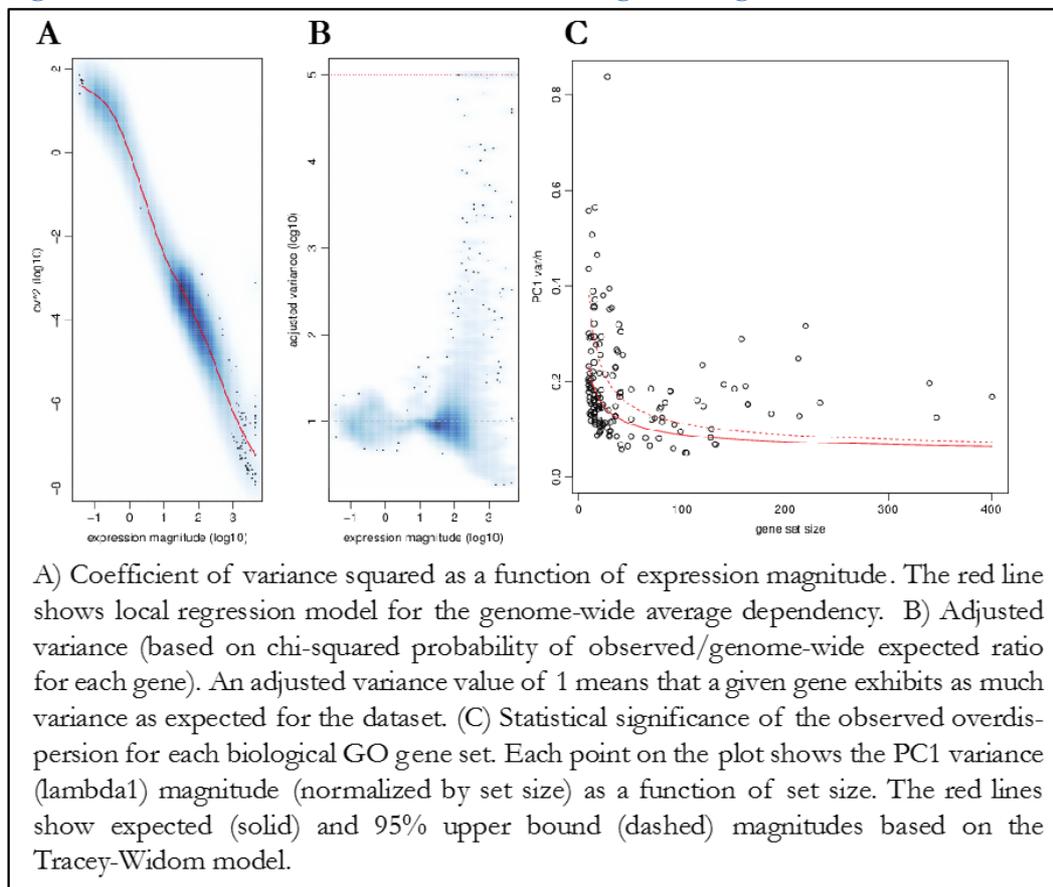
In a second step, expected levels of technical and intrinsic biological noise were normalised out of single cell error models. The variance of the NB/Poisson mixture processes, derived from the initial error modelling step, are modelled as a chi-squared distribution, using adjusted degrees of freedom and observation weights based on the drop-out probability of a given gene. Additionally, the 3 most extreme cells were trimmed and the maximum adjusted variance was limited to 5 (panel A & B, Figure 6-13).

Figure 6-12 Modelling technical error and standard biological fluctuation, to assess the extent of transcriptome variability within the 10ss anterior *scf<sup>+</sup>* population.



Genes with high-adjusted variance are ‘over-dispersed’ within the measured population and most likely show subpopulation-specific expression. To control for variation in gene coverage between samples (estimated as a number of genes with non-zero magnitude per cell) an additional normalization step is carried out (Figure 6-14). 728 genes showed significant variance between the single cell transcriptomes, following normalization for technical and intrinsic biological noise, plus differences in coverage, these genes were classed as highly variable genes (HVGs). The top 50 most over-dispersed HVG genes are indicated in table 6-1.

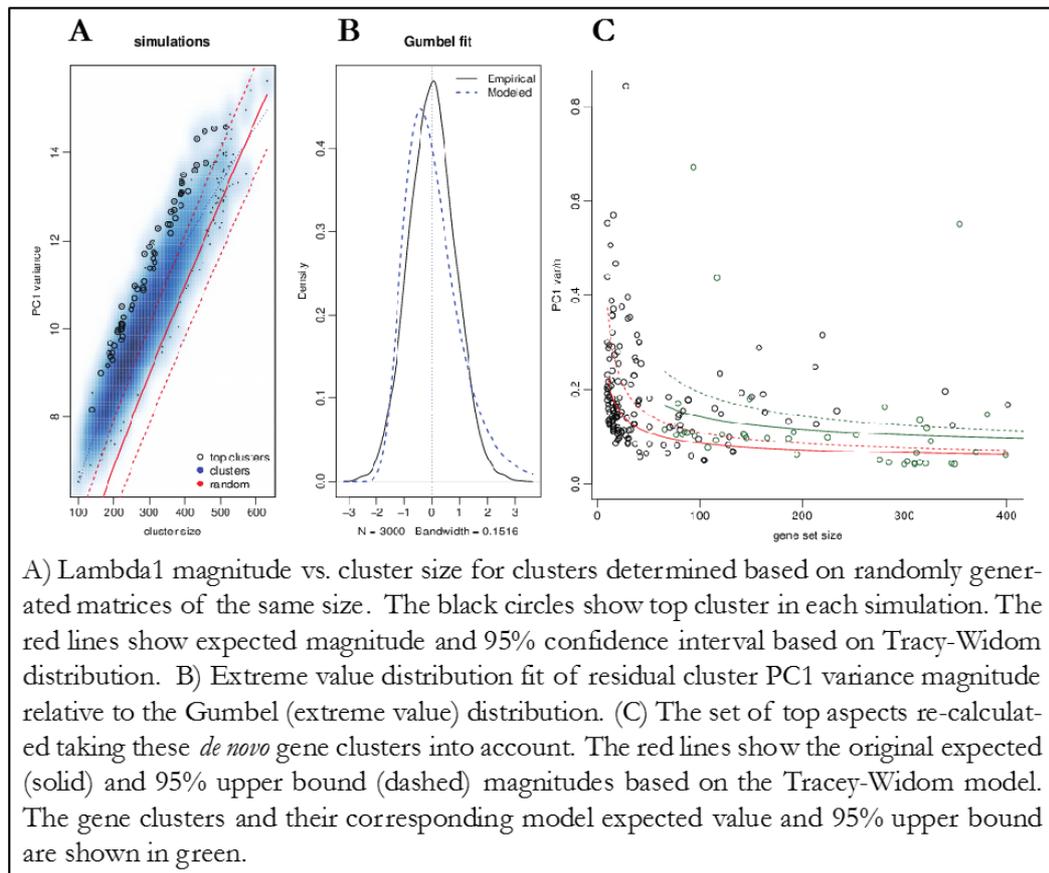
**Figure 6-13 Normalisation of variance and relating to biological GO function.**



To evaluate cellular heterogeneity, a pre-defined gene set (GO annotation obtained from the Bioconductor package `org.Dr.eg.db`) was used to identify statistically significant “excess variability”. For each gene set, if the amount of variance explained by the first principal component significantly exceed the background expectation, the gene set was judged as significant (above dashed red line in panel C, Figure 6-13).

PAGODA was then used to evaluate over-dispersion of *de novo* gene sets, and build a background model for the expectation of the gene cluster weighted principal component magnitudes (Figure 6-14).

**Figure 6-14 Background distribution of the first principal component variance magnitude.**



Cell clustering on normalized transcriptomes was carried out first taking into account all significant aspects and then based on aspects showing similar patterns (panel A, Figure 6-15). Four categories of aspects were identified. The GO terms that best describe each PC1 for each group of combined aspects were identified (panel B, Figure 6-15): Set 1: blood vessel development, Set 2: immune system process, Set 3: mitotic cell. With 3 cell clusters in green, black and red: the genes can be grouped into 4 groups of genes that are highly variable (black) labelled as 1 and less highly variable (light grey) labelled as Set 4 (Panel C, Figure 6-15).

Examination each set revealed that the first most ‘over-dispersed’ cell cluster annotation group (panel D, Figure 6-15) was associated to blood vessel development. Cells in cluster 1 (green) displayed low levels of expression for these genes, while the two other cell clusters expressed these genes at high levels. This relationship was reversed for genes from *tie1* and *fli1b*.

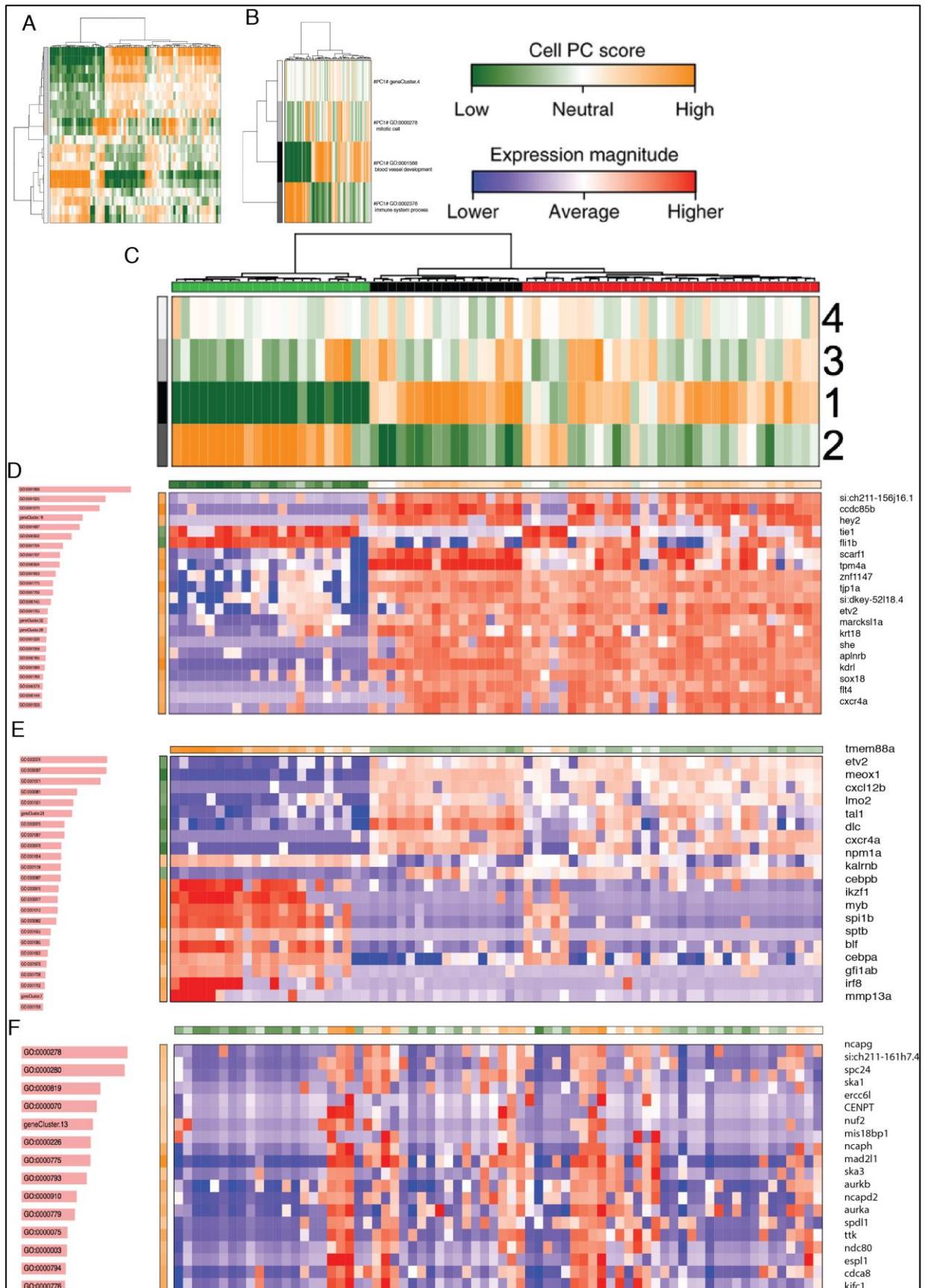
The second most ‘over-dispersed’ cell cluster (panel E, Figure 6-15) was associated with the following GO terms: immune system process, hemopoiesis and nucleic acid binding transcription factor activity. The green cell cluster cells expressed genes *tmem88a*, *etv2*, *meox1*, *cxcl12b*, *lmo2*, *scl*, *dlc* and *cxc4a* at low levels, whilst genes *cebpb*, *ikzf1*, *myb*, *spi1b*, *sptb*, *blf*, *cebpa*, *gf1ab*, *irf8* and *mmp13a* were highly expressed. This relationship was reversed in the black and red cell clusters.

The third set of genes was associated with the cell cycle, with top GO terms mitotic cell cycle, nuclear division and sister chromatid segregation. Each cell cluster contained 3-8 cells expressing genes necessary for mitosis indicating that each population was undergoing cell division. Inclusion of additional GO cell cycle-related terms (mitotic sister chromatid segregation, mitotic cell cycle, chromosome, centromeric region, condensed chromosome, condensed chromosome, centromeric region, microtubule cytoskeleton organization, kinetochore, cell cycle checkpoint, condensed chromosome kinetochore) observed earlier (Table 6-2), resulted in identical clustering, confirming that cells in S-phase were present among each cell cluster.

#### Legend for figure 6-15 (figure overleaf)

(A) Clustering of 72 single cells. Columns correspond to cells. Rows are different significant aspects, clustered by their similarity pattern. The green-to-orange colour scheme shows low-to-high weighted PCA scores (aspect patterns), where orange indicates higher expression. Blocks of colour on the left margin show which aspects have been combined by the command above. (B) Heatmap with GO terms contributing most to PC1 for each: Set1: blood vessel development, Set2: immune system process, Set3: mitotic cell, Set4: low over-dispersion *de novo* gene set “gene cluster 4” (C) Heatmap demonstrating 3 cell clusters and how the four sets of genes are expressed in each. (D) Heatmap demonstrating the 20 most over-dispersed genes corresponding to set 1- blood vessel development. (E) Heatmap demonstrating the 20 most over-dispersed genes corresponding to set 2- immune system process. (F) Heatmap demonstrating the 20 most over-dispersed genes corresponding to set 3- mitotic cell.

Figure 6-15 PAGODA analysis of single cell 10ss anterior *scf*<sup>+</sup> RNA-seq



### 6.3.3. Transcriptomic diversity of 10ss anterior *scf*<sup>+</sup> cells based on bulk RNA-seq.

This transcriptomic data was investigated for diversity in expression of the most highly expressed factors identified from the whole 10ss anterior *scf*<sup>+</sup> population transcriptomes, upon comparison to the 10ss anterior *scf*<sup>-</sup> transcriptomes. Roughly 50% of these highly expressed factors showed a stable expression level across the 72 individual *scf*<sup>+</sup> cell transcriptomes, thus do not contribute to cellular diversity at this stage. These stably expressed genes contributed to ubiquitous processes including ATP synthesis, dorsal-ventral patterning and aerobic metabolism. Mitochondrial cytochrome b, *mt-ctb*, is expressed highly in all 72 cells, and could be used as a positive control in the future.

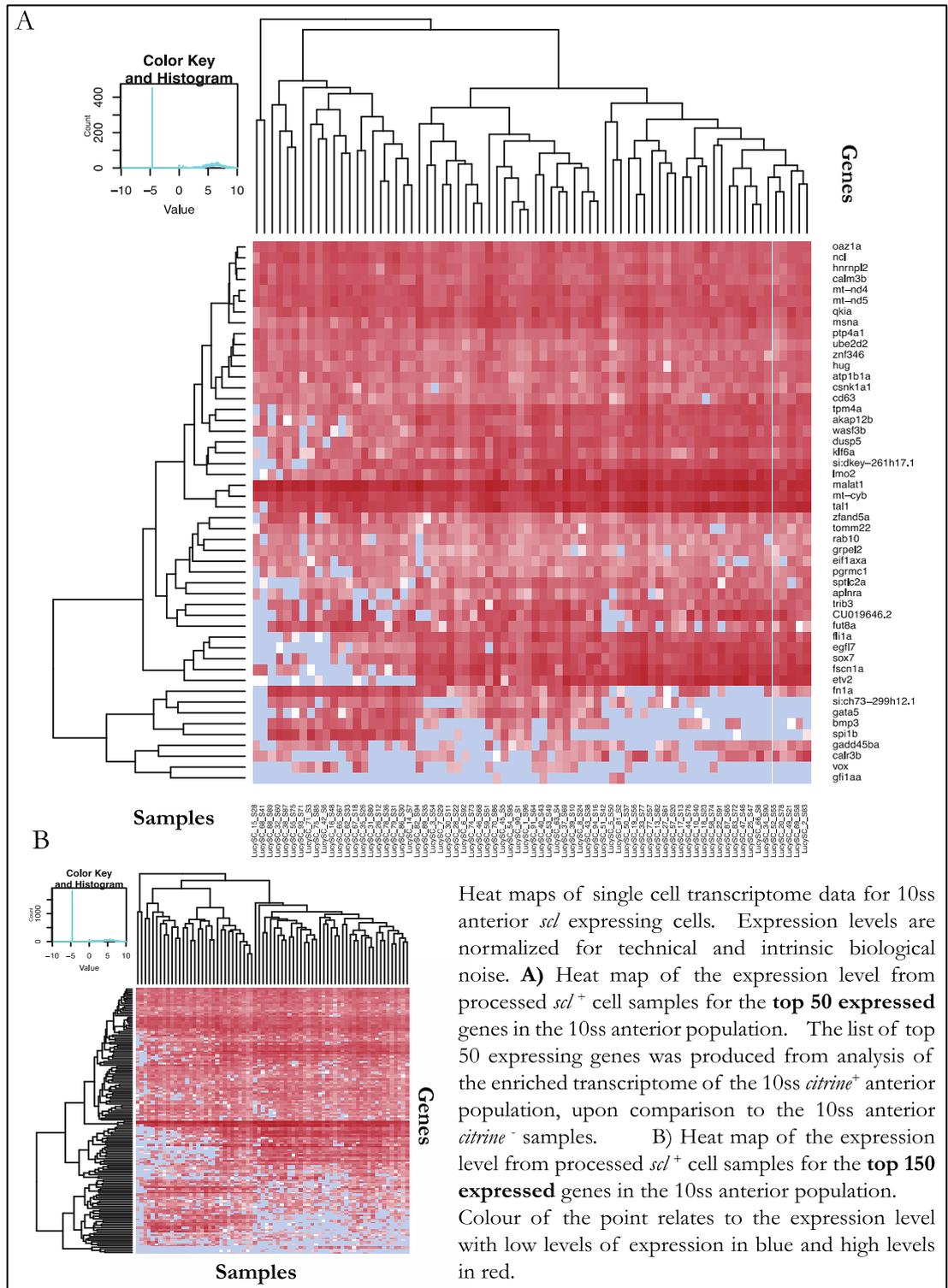
All but one cell shows high *scf* expression. In this one sample *scf* expression is 0 FPKM, while other genes such as *mt-cyb* have reads mapping to them, indicating that this is an *scf*<sup>-</sup> cell. Further investigation into the transcriptome of this single *scf*<sup>-</sup> cell revealed that no reads were detected mapping to the *citrine* sequence either. This may be the result of inaccurate FACS aliquoting a *citrine*<sup>-</sup> cell instead of a *citrine*<sup>+</sup>.

Figure 6-16 highlights groups of these highly expressed anterior genes that show differential expression across the 72 single cell transcriptomes. Two sets of genes show mostly mutually exclusive expression levels between two groups of the single cell transcriptomes. A third group of cells exist clustered by profile similarity between these two mutually exclusive cell groups, these medial clustered cells show expression of genes that are mutually exclusive in cell groups 1 and 2. These cells are highlighted yellow in figure 6-17.

The larger of the two groups of cells (group 2, figure 6-17) showing difference in profile are associated with high expression of key vascular genes including *fli1a*, *sox7* and *hey2*. The high expression of these vascular factors is combined with minimal

expression of a second set of genes, which includes key haematopoietic factors, *gfi1aa/b* and *spi1b* that are required through a range of haematopoietic lineages.

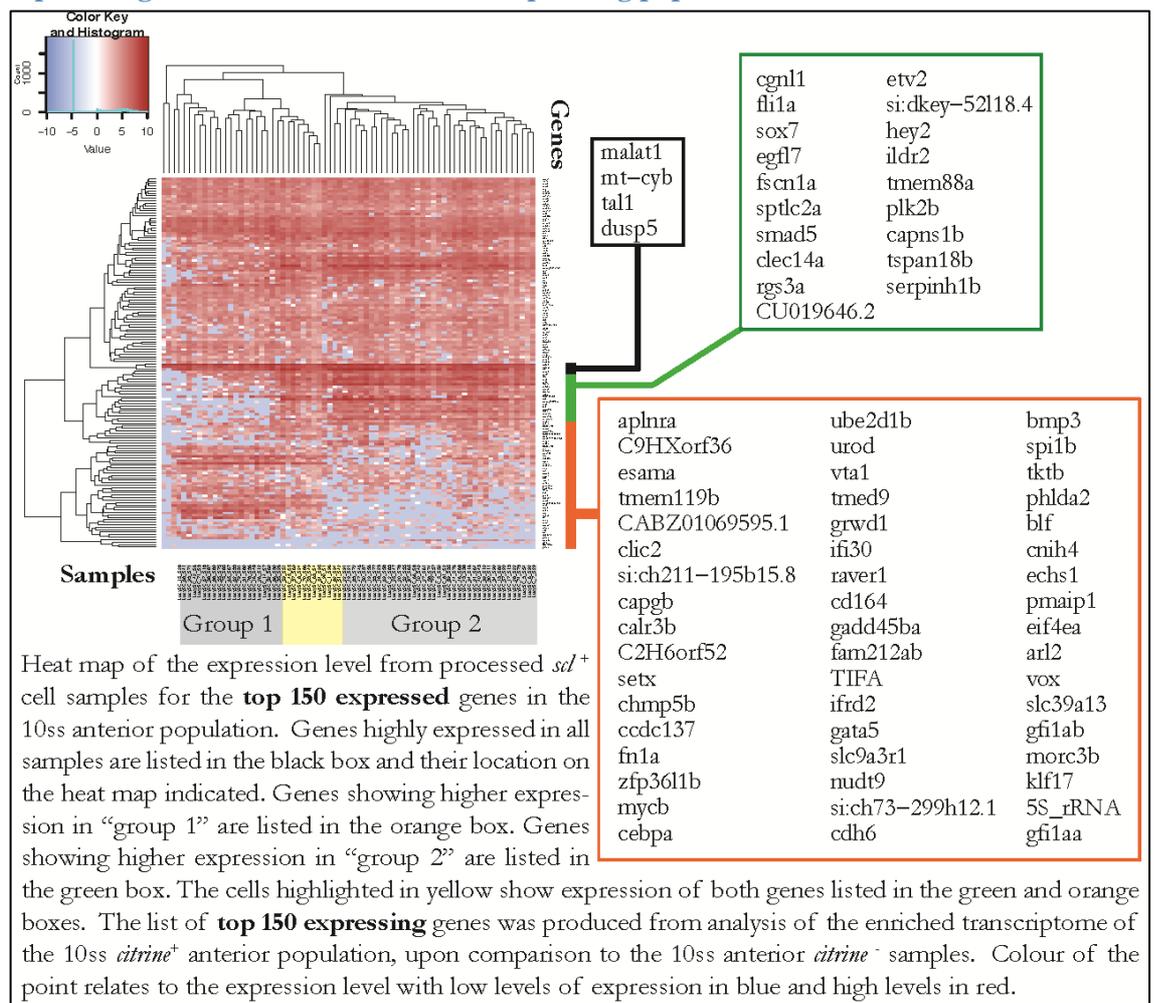
**Figure 6-16 Heat maps of expression level distribution of the top expressed genes of the 10ss anterior *scl* expressing population.**



The single cell co-expression of *spi1b* and *gfi1a/b*, which have been shown at later stages of development to drive development of opposing haematopoietic

lineages<sup>275,281-283</sup> may represent multi-lineage-priming or a very early role for these factors in haematopoietic specification. In a smaller group of cells (group 1, figure 6-17) these expression patterns are reversed with high expression of the second “haematopoietic” gene set and minimal expression of the “vascular” gene set. The cells highlighted in yellow displayed expression of genes from both haematopoietic and vascular gene sets, including *gata5*, *bmp3* and *spi1b* from the haematopoietic gene set and *etv2*, *hey2* and *tmem88a* from the vascular gene set. The existence of such cells provides further evidence for the close developmental origin of these two cell types *in vivo*.

**Figure 6-17 Annotated heat map of expression level distribution of the top 150 expressed genes of the 10ss anterior *scf* expressing population.**



These analyses further support that the 10ss anterior *scl*<sup>+</sup> population is heterogeneous, with genes for different biological functions being divided between *scl*<sup>+</sup> sub-populations.

#### 6.4. Chapter 7 summary

RNA-seq and chromatin accessibility datasets described the 10ss anterior *scl*<sup>+</sup> population as a biologically diverse and transcriptionally dynamic population, unlike the other *scl*<sup>+</sup> populations investigated in this project. This led my investigation towards further probing the nature of this diverse early anterior haematopoietic and vascular progenitor population. Crosses of transgenic zebrafish lines expressing fluorescent reporter genes under the control of the lateral plate mesoderm marker, *gata5* or the key vascular developmental receptor, *kdrl*, with my Tg(*scl*BAC:*scl*-flag-*tevi-2A-citrine*)line were used to assess the interaction and overlap of these populations *in vivo*.

*gata5*<sup>+</sup> and *scl*<sup>+</sup> populations showed partial overlap at the 10ss. By the 20ss overlap between *gata5*<sup>+</sup> cells and *scl*<sup>+</sup> cells varied with the *scl*<sup>+</sup> subpopulation; *scl*<sup>+</sup> cells poised to become cranial vasculature lacked *gata5* expression, as did *scl*<sup>+</sup> at the very centre of the heart field. However these central *scl*<sup>+</sup> *gata5*<sup>-</sup> cardiac cells were surrounded by *gata5*<sup>+</sup> cells, some of which were also *scl*<sup>+</sup>. *scl*<sup>+</sup> cells on top of the yolk were also *gata5*<sup>+</sup>, indicating that the co-expression of *gata5* and *scl* may favour the myeloid lineage over the vascular fate.

Expression of *kdrl* at the 10ss showed varying degrees of overlap with the *scl* population, indicating that this may be a dynamic stage in endothelial cell specification. By the 20ss the *kdrl* and *scl* populations in the anterior of the embryo are highly overlapping, with co-expression of the two reporters in the central *scl*<sup>+</sup> population in the heart field (putative endocardium) and at lower levels in the

progenitors of the cranial vasculature. Cells expressing *scf* over the yolk however do not express *kdrl*, further supporting the conclusion that these cells have adopted the myeloid fate.

Single cell RNA-seq was carried out to further resolve the transcriptional diversity of the 10ss anterior *scf*<sup>+</sup> population. Over 700 genes were identified as being significantly highly variably expressed across the single cell transcriptomes, after accounting for biological and technical noise. These highly variable genes three different biological profiles within the 10ss anterior *scf*<sup>+</sup> population; vascular, haematopoietic and both vascular and haematopoietic. The latter of these cell groups may represent the proposed haemangioblast.

Table 6-1 Top 50 most over-dispersed genes in single cell RNA-seq datasets

Ensembl Gene ID	Gene Name	Variance
ENSDARG00000000767	spi1b	5
ENSDARG00000006019	tktb	5
ENSDARG00000012395	mmp13a	5
ENSDARG00000013539	ikzf1	5
ENSDARG00000014522	cdh6	5
<b>ENSDARG00000019930</b>	<b>tal1</b>	<b>5</b>
ENSDARG00000020850	eef1a1l1	5
ENSDARG00000021443	zfp3611b	5
ENSDARG00000036074	cebpa	5
ENSDARG00000037870	actb2	5
ENSDARG00000042725	cebpb	5
ENSDARG00000043126	blf	5
ENSDARG00000056407	irf8	5
ENSDARG00000058160	tnfaip2b	5
ENSDARG00000063905	mt-co1	5
ENSDARG00000077473	mych	5
ENSDARG00000080337	AC024175.4	5
ENSDARG00000091234	CU019646.2	5
ENSDARG00000099970	malat1	5
ENSDARG00000102808	calr3b	5
ENSDARG00000104474	il6r	5
ENSDARG00000019815	fn1a	4.977265
ENSDARG00000104370	esm1	4.772521
ENSDARG00000068992	hspa8	4.749606
ENSDARG00000028335	hmga1a	4.674238
ENSDARG00000077341	ppp1r14c	4.653853
ENSDARG00000053666	myb	4.283529
ENSDARG00000018404	krt18	4.238393
ENSDARG00000057633	cxcr4a	4.136026
ENSDARG00000103720	ZFP36	4.080769
ENSDARG00000037879	lfng	4.068734
ENSDARG00000038066	kpna2	3.958882
ENSDARG00000019130	plk2b	3.792321
ENSDARG00000013576	gadd45bb	3.775715
ENSDARG00000007823	atf3	3.726259
ENSDARG00000019949	serpinh1b	3.669938
ENSDARG00000053868	etv2	3.618613
ENSDARG00000017821	gata5	3.467021
ENSDARG00000020133	jdp2b	3.463762
ENSDARG00000070670	crip2	3.439666
ENSDARG00000015449	fut8a	3.403121
ENSDARG00000075549	cdh5	3.400832
ENSDARG00000055723	hsp70l	3.373204
ENSDARG00000061848	col15a1b	3.373109
ENSDARG00000100657	egfl7	3.372616
ENSDARG00000001452	adam8a	3.363435
ENSDARG00000092035	si:ch211-156j16.1	3.342485
ENSDARG00000095019	lmo2	3.308579
ENSDARG00000099411	zgc:158343	3.194702
ENSDARG00000006766	snd1	3.127586

**Table 6-2 Variance of gene lists produced through GO analysis or from *de novo* evaluation of over-dispersion.**

go_name	go_term	npc	n	score	z	adj.z
vasculogenesis	GO:0001570	1	28	6.084370694	25.44479885	25.31513976
NA	geneCluster.16	1	363	5.445286611	19.45134294	19.24971237
blood vessel development	GO:0001568	1	220	4.485242442	36.39694717	36.25700296
angiogenesis	GO:0001525	1	158	3.793386586	27.5402071	27.40565949
NA	geneCluster.25	1	441	3.514828685	15.42908247	15.21991803
hemopoiesis	GO:0030097	1	213	3.509503384	28.01146836	27.86888179
mesoderm formation	GO:0001707	1	16	3.199309213	11.14360573	10.96022194
immune system process	GO:0002376	1	340	3.058507511	28.19408227	28.03798233
NA	geneCluster.13	1	144	3.026554791	9.88638446	9.600586906
formation of primary germ layer	GO:0001704	1	30	2.970928059	12.34477158	12.15176541
ameboidal-type cell migration	GO:0001667	1	120	2.855774556	18.32789795	18.15618862
peptide receptor activity	GO:0001653	1	18	2.76239136	9.62287249	9.436862746
sister chromatid segregation	GO:0000819	1	32	2.727879884	11.32054965	11.1262969
nucleic acid binding transcription factor activity	GO:0001071	1	401	2.697097052	25.72966818	25.59433576
mitotic sister chromatid segregation	GO:0000070	1	30	2.636917978	10.61319278	10.42928626
RNA polymerase II core promoter proximal region sequence-specific DNA binding	GO:0000978	1	39	2.618626687	11.42621935	11.22335789
transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding	GO:0000982	1	24	2.545283517	9.432428871	9.25031133
core promoter proximal region DNA binding	GO:0001159	1	41	2.544526596	11.17488286	10.98760892
core promoter proximal region sequence-specific DNA binding	GO:0000987	1	41	2.544526592	11.17488284	10.98760892
exocyst	GO:0000145	1	13	2.543018259	7.772590967	7.57240031
RNA polymerase II transcription factor activity, sequence-specific DNA binding	GO:0000981	1	162	2.51461343	17.27764819	17.10324452
nuclear division	GO:0000280	1	141	2.469244303	16.09870256	15.9254412
RNA polymerase II transcription factor binding	GO:0001085	1	10	2.461229388	6.822516403	6.603129586
skeletal system development	GO:0001501	1	151	2.402409538	15.88053781	15.71093203
mitotic cell cycle	GO:0000278	1	234	2.233805669	16.67240595	16.49876257
chromosome, centromeric region	GO:0000775	1	43	2.21874548	9.332606555	9.152208512
cell activation	GO:0001775	1	37	2.169506318	8.578840303	8.390237334
condensed chromosome	GO:0000793	1	38	2.16225374	8.609187065	8.417526162
condensed chromosome, centromeric region	GO:0000779	1	16	2.086443215	6.091809874	5.859686971
response to yeast	GO:0001878	1	14	2.033583188	5.560807886	5.324615841
regulatory region DNA binding	GO:0000975	1	164	2.01129743	12.5881383	12.39334273
regulatory region nucleic acid binding	GO:0001067	1	164	2.011297402	12.58813801	12.39334273
RNA polymerase II regulatory region sequence-specific DNA binding	GO:0000977	1	89	1.984141543	9.973927591	9.790638567
RNA polymerase II regulatory region DNA binding	GO:0001012	1	89	1.98414154	9.973927571	9.790638567
patterning of blood vessels	GO:0001569	1	15	1.946829428	5.242292546	5.000797292
cell morphogenesis	GO:0000902	1	347	1.939688145	15.4348558	15.26601486
transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding	GO:0001228	1	22	1.929550334	5.884470637	5.652877178

transcription regulatory region sequence-specific DNA binding	GO:0000976	1	115	1.925910009	10.360119	10.17586459
morphogenesis of a polarized epithelium	GO:0001738	1	15	1.918585247	5.092508734	4.848429219
embryonic axis specification	GO:0000578	1	10	1.909829695	4.363986731	4.10422208
somitogenesis	GO:0001756	1	70	1.906736721	8.574402614	8.389321143
NA	geneCluster.32	1	152	1.868972467	6.357605491	5.957482155
cytokinesis	GO:0000910	1	36	1.863692347	6.52907085	6.304162455
neural crest cell migration	GO:0001755	1	36	1.858075683	6.490970122	6.268909611
microtubule cytoskeleton organization	GO:0000226	1	121	1.812458259	9.462811958	9.277538445
cell morphogenesis involved in differentiation	GO:0000904	1	187	1.811206192	11.02165442	10.84049683
eye development	GO:0001654	1	214	1.792794506	11.34743338	11.14855611
morphogenesis of a branching structure	GO:0001763	1	20	1.754888885	4.664011265	4.407314365
condensed nuclear chromosome	GO:0000794	1	15	1.749975732	4.16084204	3.893565591
urogenital system development	GO:0001655	1	84	1.703004843	7.346423899	7.138684288
regulation of cell growth	GO:0001558	1	51	1.66528229	5.832779686	5.603341211
NA	geneCluster.28	1	229	1.65857292	6.185418596	5.811439572
ossification	GO:0001503	1	21	1.658310368	4.135286571	3.875619911
gastrulation with mouth forming second	GO:0001702	1	33	1.650969465	4.85968302	4.608669185
endoderm formation	GO:0001706	1	17	1.59591215	4.029176912	3.772036745
rRNA modification	GO:0000154	1	15	1.583920855	3.799782381	3.537126286
kidney development	GO:0001822	1	80	1.580329665	6.081809142	5.853519568
MAPK cascade	GO:0000165	1	79	1.525743672	5.52437313	5.290857048
kinetochore	GO:0000776	1	26	1.512474107	4.058570473	3.798747816
SNARE binding	GO:0000149	1	41	1.509945003	4.138398322	3.875619911
cell cycle checkpoint	GO:0000075	1	40	1.483599237	4.537645543	4.278601389
condensed chromosome kinetochore	GO:0000777	1	12	1.419502385	2.652773245	2.328285284
G-protein coupled receptor binding	GO:0001664	1	40	1.415010586	3.898978275	3.63820414
neuron migration	GO:0001764	1	20	1.361214989	2.742681036	2.422234357
reproduction	GO:0000003	1	81	1.355853201	4.409109539	4.147380623
positive regulation of protein phosphorylation	GO:0001934	1	74	1.301760019	3.659729296	3.394758261
NA	geneCluster.7	1	311	1.288824083	4.251636496	3.788464076
liver development	GO:0001889	1	71	1.252270991	3.087048525	2.789306567
chromatin	GO:0000785	1	93	1.250702243	3.381552339	3.1032329
regulation of protein phosphorylation	GO:0001932	1	129	1.239267795	3.657784783	3.394758261
negative regulation of transcription from RNA polymerase II promoter	GO:0000122	1	81	1.237984172	3.075374712	2.781782478
NA	geneCluster.4	1	686	1.231619774	4.951336103	4.521436452
NA	geneCluster.37	1	158	1.210503723	2.708801921	2.080117021
NA	geneCluster.31	1	528	1.171600045	3.751075043	3.263490842

The z column in Table 6-2 gives the Z-score of pathway over-dispersion relative to the genome-wide model (Z-score of 1.96 corresponds to P-value of 5%, etc.). "z.adj" column shows the Z-score adjusted for multiple hypothesis (using Benjamin-Hochberg correction). "score" gives observed/expected variance ratio "sh.z" and "adj.sh.z" columns give the raw and adjusted Z-scores of "pathway cohesion", which compares the observed PC1 magnitude to the magnitudes obtained when the observations for each gene are randomized with respect to cells. When such Z-score is tight when multiple genes within the pathway contribute to the coordinated pattern.

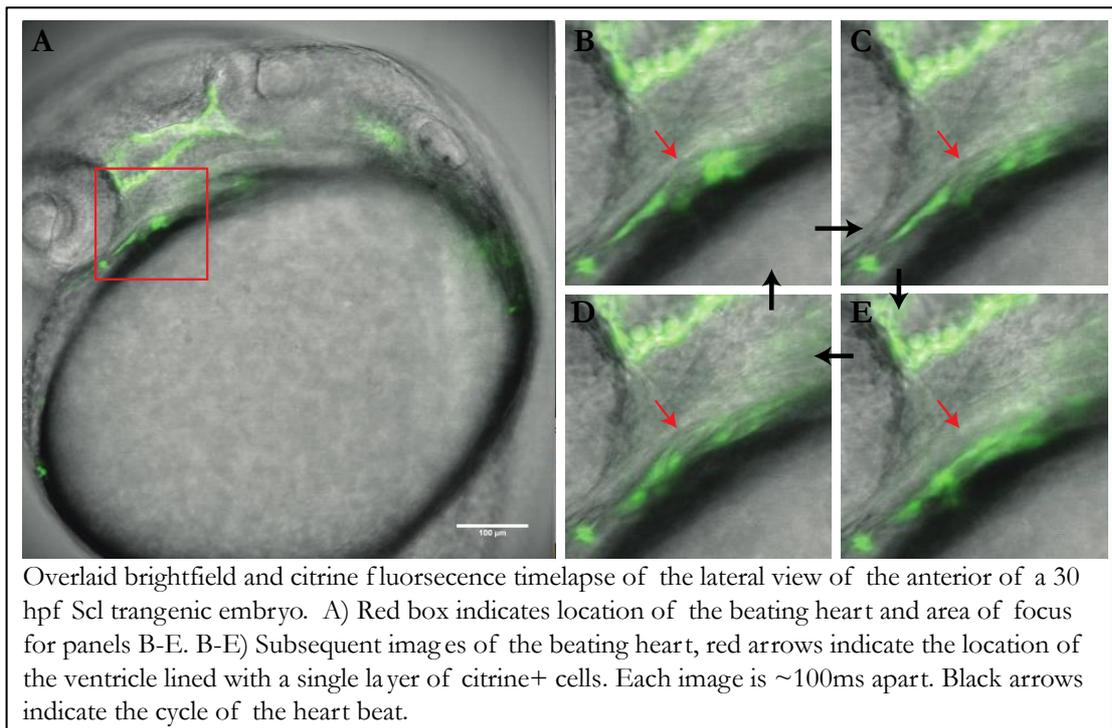
## 7. Discussion

During the course of this project I have generated, characterised and used a transgenic zebrafish line to investigate the nature of primitive blood and vascular cells, known to express the master regulator *scf* during early vertebrate development.

I used BAC transgenesis to generate a tagged-*scf* and reporter gene construct, which through Tol2-mediated transposition was successfully incorporated into the genome of zebrafish in a heritable fashion. The *citrine* reporter gene expressed in this transgenic line was shown to closely mimic endogenous *scf* expression patterns, and was used to identify and specifically isolate *scf*<sup>+</sup> cells by FACS analysis.

High resolution imaging of the *scf*<sup>+</sup> population *in vivo* revealed the range of cell migration and proliferation events that these cells undergo in early development. The posterior *scf*<sup>+</sup> population initiates first as a pair of bilateral stripes, which move medially in a “zipper-like” movement and merge at the midline to form the ICM. The anterior *scf*<sup>+</sup> population showed greater motility and *citrine*<sup>+</sup> cells were observed to contribute to a range of cell types by the 20ss stage. From an initial anterior bilateral small *scf*<sup>+</sup> population, cells moved anteriorly and formed cranial vasculature, laterally over the yolk to adopt the myelopoietic lineage and medially to potentially contribute to the developing endocardium. This correlates with previous studies of *scf* expressing cells in zebrafish embryos<sup>17,104</sup>, however I have resolved a previously unreported contribution to the endocardium from the anterior *scf*<sup>+</sup> population (Figure 7-1). Using a mutant zebrafish line *scf* expression has been shown to be required for the proper migration of endocardial precursors into the heart field, though this study did not identify in which cells expression was necessary<sup>47</sup>.

Figure 7-1 Time-lapse images of the beating heart from a 30hpf *scf*-transgenic embryo



In the zebrafish line generated in my project, strong expression of the citrine reporter gene is driven in cells migrating into the heart field and in a single layer of centrally localized cells, within the beating heart. This line could be useful for further studies on cardiac formation and defects at early stages of development, before the expression of the endocardial marker *nfatc1* is detectable<sup>394</sup>.

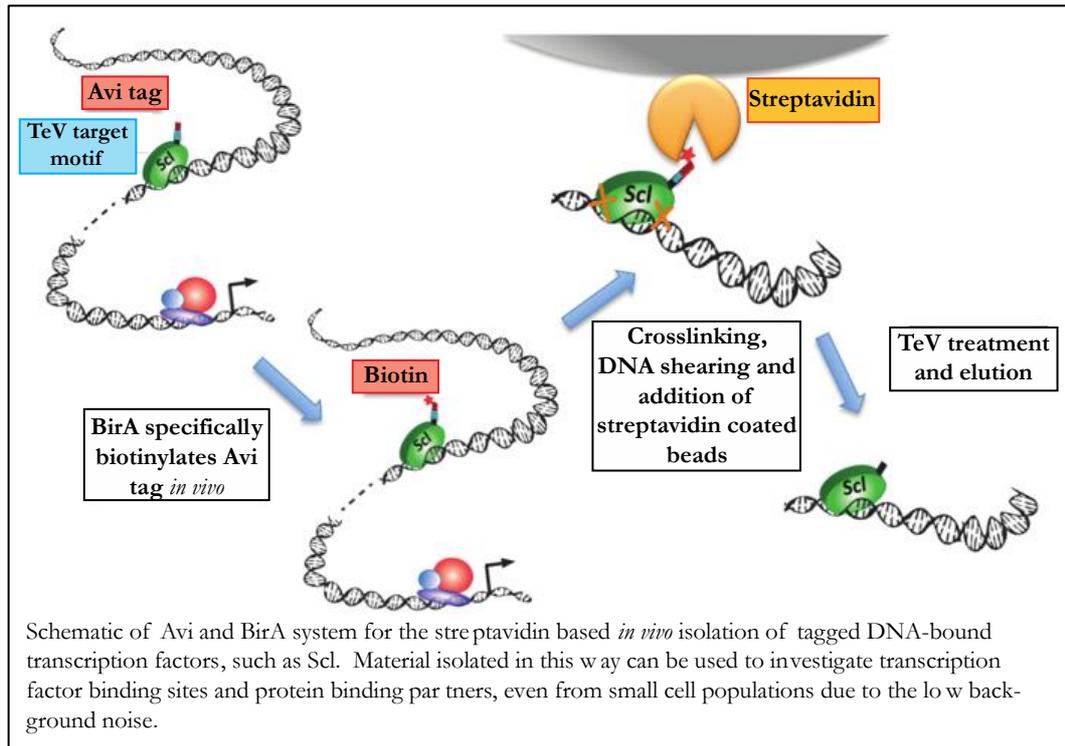
The transgenic line generated in this project recapitulates endogenous expression patterns and those observed in other *scf* transgenic lines<sup>198,222,223</sup>. Rodaway *et al.* only label the *scf- $\alpha$*  isoform<sup>198</sup>, while the papers from Zilong's lab describe zebrafish lines with each of the isoforms labeled with different fluorophores<sup>222,223</sup>. My transgenic line labels both isoforms simultaneously, and I have shown that it can be used for tracing the movements of individual cells in a similar fashion to previous studies<sup>198</sup>. I would ideally use my line in combination with other early mesodermal markers to trace the origins of the hematopoietic *scf* expressing progenitors.

The transgenic line generated in this project not only permits the visualisation of *scf*<sup>+</sup> cells through early vertebrate development but also permits the isolation by FACS and the purification of Scl protein, through use of biotin-streptavidin pull-down, all from a true *in vivo* source.

*In vivo scf*-expressing cells are present in small numbers, intermingled with other embryonic cell types and no specific zebrafish Scl antibody is available that could sustain stringency of mini-ChIP procedures, required for background-free amplification of isolated chromatin. Thus this transgenic *scf* coding sequence was designed to include an avi tag, which could be used to exploit the strength of the biotin:streptavidin interaction for efficient Scl protein isolation. Subsequent stringent washes are possible with this system and would permit the collection of Scl:DNA complexes at a sufficient purity for sequencing to provide genome-wide Scl binding maps, even from these small transient cell populations.

ChIP-seq has been previously carried out on this master regulator at the genome level, however these have been in cell culture or primary culture<sup>126,131,395,396</sup>. The transgenic system developed in this project would enable Scl:DNA complex isolation from a true *in vivo* source (Figure 7-2), which have spatial and temporal specificity, and can be directly combined with transcriptomic and chromatin accessibility maps for the same cell populations.

Figure 7-2 Schematic of BirA mediated biotin chromatin precipitation.

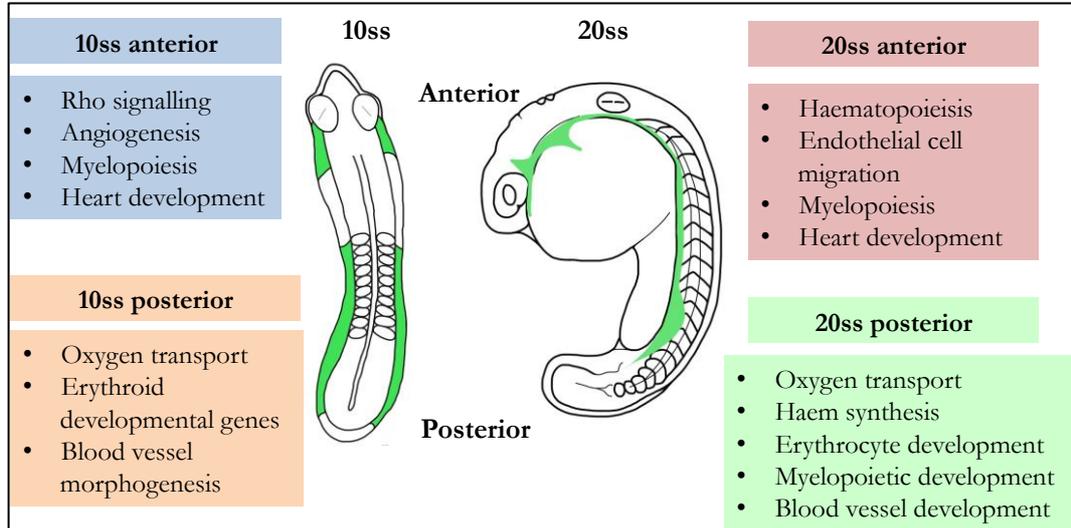


The use of differential expression analysis between cellular contexts has been highly valuable towards describing each of the *scf* expressing contexts. Transcriptome analysis alone offers an insight into the cell type of interest, but without comparison to other cells, expression levels cannot be put into biological context. My analysis of gene enrichment between *scf<sup>+</sup>* and *scf<sup>-</sup>* cells from the same embryonic origin has identified a significant number of novel genes that show specific enrichment within *scf<sup>+</sup>* populations. These novel factors would be interesting to validate their enrichment in haematopoietic and vascular progenitors *in vivo* and to investigate further as they may contain key regulators of early developmental decisions.

Enriched transcriptome analysis has provided a prediction of biological function for each of the contexts studied at a genome wide resolution. Despite the reliance on published studies for annotation, PANTHER analysis identified significantly over representation of specific gene classes associated with each of the *scf<sup>+</sup>* context's enriched gene sets (Figure 7-3). The over-represented biological functions identified

through this approach correlate with previous conclusions drawn from single gene investigations<sup>205,278,294,389</sup>.

**Figure 7-3 Summary of top enriched gene ontology terms for each *scf*<sup>+</sup> population investigated in this project**



Based on my transcriptomic data I propose that the posterior *scf*<sup>+</sup> population is developmentally further progressed than the anterior, matching the onset of *scf* expression in the posterior at 3ss and in the anterior at 6ss<sup>17</sup>. A strong erythroid profile was already adopted by the 10ss posterior *scf*<sup>+</sup> population, while the anterior showed enrichment on genes relating to a range of biological processes, without a strong enrichment for any specific function.

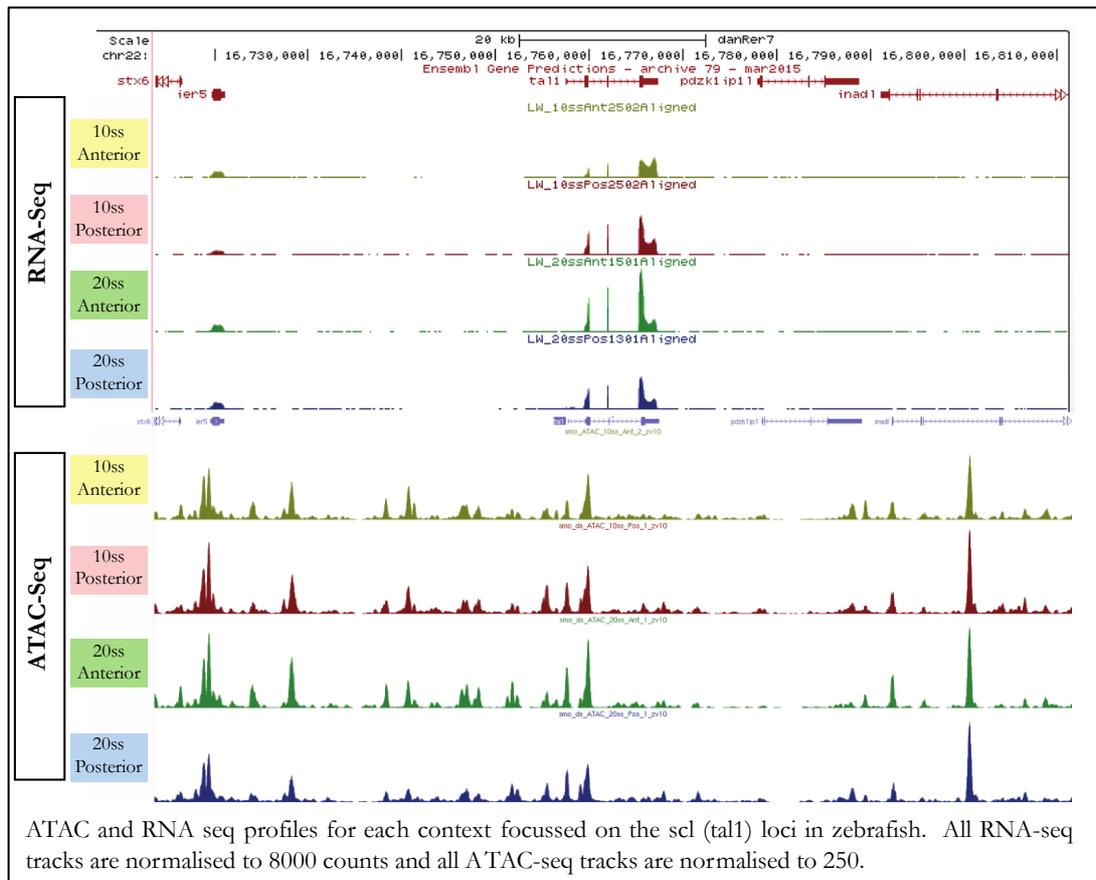
ATAC seq supported the theory that the posterior *scf*<sup>+</sup> was further differentiated than the anterior counterpart with 90% of the open promoter peaks correlating with detectable transcription of the gene plus the maintenance of this expression through to the 20ss. In contrast only 60% of the 10ss anterior promoter associated chromatin peaks, displayed significant expression of the gene, thus suggesting that these genes were poised for transcription, but had yet to commence. In further analysis I would like to investigate the identity and function of these “poised” genes and determine whether they do become expressed and contribute to fate decisions at a later stage of development.

The open chromatin maps generated in this project not only identify potentially “poised” genes, but also putative regulatory elements and their context specific behaviour. Initially binding site analysis guided by my transcriptome data can be used to propose factors that may bind and regulate target genes, and thus begin to build a potential gene regulatory network (GRN). In combination with transcription factor binding maps from ChIP-seq experiments, novel direct regulatory interactions can be discovered and GRN circuits can be tested.

GRNs have been proposed for several developmental processes including mesoderm formation and specification of the erythroid lineage<sup>397-402</sup>. These networks provide comprehensive models of current knowledge of regulatory interactions, which can provide insight into common processes in diverse systems and predict novel interactions. Expression of the *scf* gene is involved in the development of multiple key cell types and drives leukemic transformation in many cases of childhood leukemia, thus development of a GRN surrounding *scf* could be used to inform a range of academically and clinically important research. It would be of great interest to investigate if a common program was used in anterior and posterior haematopoietic and vascular progenitors, plus in neural development, that is elaborated upon to provide tissue-specific effects.

My open chromatin maps offer a genome wide tool to identify potential novel regulatory regions including within the *scf* loci (Figure 7-4 shows these maps). Regulation of the *scf* loci itself is a matter of great interest due to its contribution to T-ALL and the key role *scf* plays in development and has been highly studied. These datasets offer the possibility to identify novel putative *scf* regulatory regions and relate their presence to a specific biological context.

Figure 7-4 ATAC and RNA seq reads mapped to the zebrafish *scf*/loci, for each of the contexts studied in this project.



Previous studies have identified regulatory regions within the 3' UTR of *scf* in a range of vertebrate studies<sup>94,155</sup>. In each of my datasets a region within the 3'UTR shows high levels of transcription in close proximity to an area of open chromatin. Further sequence analysis could reveal whether this region is conserved with those reported in previous studies, and what transcription factor binding sites lie within this sequence. With further analysis, I may identify additional putative regulatory regions and be able to relate these to the biological context from which they arise. 4C (Chromosome Conformation Capture on Chip)<sup>403</sup> experiments could be used to assess the regulatory regions that are brought into proximity to the *scf* promoter *in vivo*. Combination of such a dataset with my maps of chromatin accessibility could be used to relate a comprehensive set of *scf* regulatory regions to biological activity, *in vivo*.

This project has profiled four early *scf* expressing contexts *in vivo*, followed by a focussed investigation on the diversity within the 10ss anterior. Ideally I would use the tools and techniques already optimised to perform a similar analysis of the posterior of the embryo. Of particular interest would be to use imaging and profiling techniques to investigate the initiation of the posterior haematopoietic program and compare this to the current anterior dataset. The posterior of the zebrafish embryo is the site of HSC development, contained within the *scf* expressing population, thus further investigation into the development of the hemogenic endothelium would be of great clinical and academic interest.

In mouse mutant studies *Scf* is not expressed in the absence of *Flk1* (mouse ortholog of zebrafish *kdrl*, whose homolog *kdrl* is absent from mice), suggesting that the action of this endothelial receptor acts upstream of *Scf* expression in early development<sup>404</sup>. In my 10ss images of double positive embryos from *Tg(kdrlBAC:gfp)* and *Tg(scfBAC:scf-flag-tev-avi-2A-citrine)* line crosses, cells positive for each of these factors show a range of overlap. This suggests that an initial wave of *kdrl* expression had previously occurred and subsequently reduced to result in the *scf*<sup>+</sup> *kdrl*<sup>-</sup> cells observed. This suggests that despite being homologs and its ortholog being absent from the mouse genome, *kdrl* activity in zebrafish haematopoiesis is significantly different to *Flk1* activity in the mouse. To confirm this in the zebrafish, imaging at a cellular resolution would be required for both *kdrl*<sup>+</sup> and *kdrl*<sup>-</sup> populations in combination with knockout studies.

Hobson *et al.* have previously shown that endothelial cells rapidly proliferate in response to VEGF signalling. This can be correlated to my hyperspectral images of the *kdrl* reporter line crossed to my transgenic *scf* line; which shows co-expression of *scf* and *kdrl* reporters in putative cranial vasculature and its progenitors. As these *scf*<sup>+</sup>

cells extend into the cranial region, *kdrl* reporter gene expression level dropped, suggesting that they reduce in proliferative capacity as these endothelial cells migrate to their final embryonic position. Ideally I would like to carry out a time-lapse using these lines to clarify the early (pre 10ss) relationship between these two populations in zebrafish, and at later time points (10-22ss) to assess the dynamic expression of *kdrl* within *scl*<sup>+</sup> cells as they adopt endothelial or endocardial fates.

Imaging of the Tg(*gata5*BAC:*la-GFP*) reporter line supports previous studies that the *scl*<sup>+</sup> population does not contribute to the myocardium<sup>286</sup> since *scl*<sup>+</sup> *gata5*<sup>-</sup> cells within the heart field are restricted to the central putative endocardial population. Instead, this central population is surrounded by *scl*<sup>-</sup> *gata5*<sup>+</sup> cells, which I propose to be the developing myocardium, based on studies in *Xenopus* with *gata4/5/6*<sup>287</sup>.

My single cell transcriptome data confirms the existence of cell population co-expressing key endothelial and haematopoietic factors. Such a cell has been proposed to be a common progenitor, the haemangioblast<sup>8,25,405,406</sup>, to both of these lineages (Figure 7-5). Contributing to this belief in a common progenitor is the co-expression of other key vascular and haematopoietic factors such as *Flk1*, *Runx1* and *GATA2*<sup>406</sup> plus the close physical origins of these lineages *in vivo*.

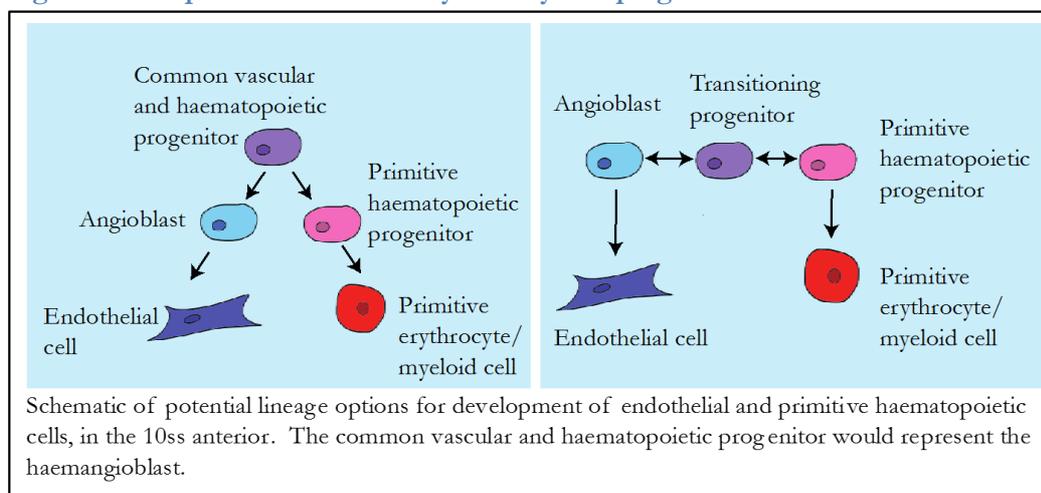
My single cell data is the first data to confirm that this co-expression is occurring *in vivo*, in individual cells rather than a population of cells. The imaging datasets with *gata5* and *kdrl* reporter lines demonstrates that these co-expressing cells exist in a small locale in proximity to the anterior lateral plate mesoderm.

An alternative explanation for the co-expression of vascular and haematopoietic genes within early single cells, may be that the population is in transition between two cellular profiles (Figure 7-5). Definitive HSC arise from the hemogenic endothelium of the dorsal aorta by endothelial to haematopoietic transition

(EHT)<sup>223,407</sup>. A similar process may occur at the initiation of primitive haematopoiesis, as a proportion of early angioblasts undergo transition into haematopoietic precursors.

It is important to consider that my single cell data offers a snapshot of the transcriptional profiles and cannot what lineages will develop from each cell group or their origins. Such lineage knowledge is required to resolve the relationship between the three cell types identified here.

**Figure 7-5 Proposed cellular activity of early *scf*<sup>+</sup> progenitor cells**



In addition to identifying potential anterior subpopulations within the *scf*<sup>+</sup> population, I have described their transcriptome and identified key variable genes that could potentially guide the lineage decisions between these cell fates.

To further understand these early fate decisions that generate the haematopoietic and vascular lineages I propose to investigate earlier stages of development, through further profiling and fluorescent *in situ* techniques. Lineage tracing using photo-convertible fluorophores could be used to study the behaviour of these early progenitor subpopulations and assess their contribution to different lineages as development progresses. Such techniques have been used previously and have the advantage of being a truly *in vivo* investigation without physical or chemical manipulation<sup>21</sup>.

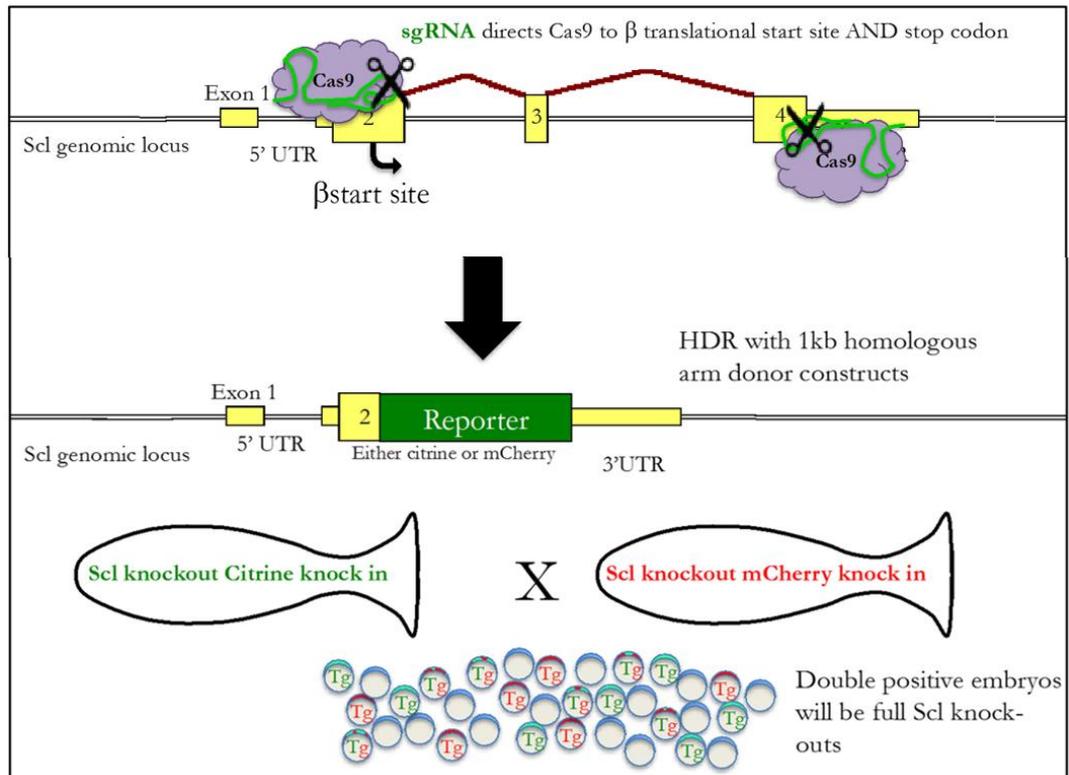
Hybridization chain reaction (HCR) *in situ* would ideally be used to identify the key populations that contribute and surround these early haemangiogenic progenitors, at a sensitivity and resolution previously unattainable. Ideally these techniques would enable the identification and isolation of the initial haematopoietic and vascular progenitor cells, so that transcriptional and chromatin accessibility profiling can be performed and used to describe these crucial cells. Knowledge of the transcriptome of these cells could inform efforts to develop induced haematopoietic progenitor cells. Recently significant developments have been made with using key transcription factors (including *scf*) to direct differentiation of human pluripotent stem cells (hPSC) towards a hemogenic endothelium and blood lineage<sup>408</sup>. Another key achievement has been the use of enforced expression of *Scf*, *Lmo1* and *Notch1* to transform thymocytes into stem cell cells<sup>409</sup>. Together these findings, based on knowledge of regulatory networks, bring the field closer to the possibility of using regenerative techniques to treat blood based disease and disorders. The ability to produce stem like cells, in the future would enable patient specific engraftments to be produced. Due to the multipotent nature of progenitor cells, such induced stem cells could be used to treat a range of ailments, including blood loss, through directed development down the erythroid lineage.

Profiles from early developmental progenitors, as described in this project, could be compared to transformed cells from T-ALL, to identify potential commonalities between developmental stem like cells and leukemogenic stem cells. Large scale studies of patient and T-ALL derived cell lines have been carried out and key abnormal features identified within their transcriptomes<sup>410</sup>. Comparison of these features may identify factors that in development maintain stemness, while in culture

or in T-ALL patients drive uncontrolled proliferation, and thus would be a key drug target.

To assess the *in vivo* contribution of *scf* to the development of the haematopoietic and vascular systems, knockout, knockdown and overexpression experiments have been previously carried out. Each of these suffer from the same inherent paradox, that it is impossible to study a factor's activity *in vivo*, if it has been effectively removed. With *scf* an added complication is its crucial nature for the survival of mouse embryos. In zebrafish, morpholino mediated knockdown of *scf* has previously been optimised and embryos survive to at least 48hpf, with undetectable levels of *scf* expression. The onset of genome engineering offers the potential for specific and efficient gene knockdown *in vivo*. This technology also offers the possibility of mediating transgenic insertion at specific genetic loci. Currently I am working on developing heterozygotic transgenic zebrafish lines in which the endogenous *scf* locus is disrupted and replaced by a fluorescent reporter gene cassette. Upon in-crossing of these lines, double positive embryos would have full knockout of *scf*, combined with fluorescent reporter expression under the control of *scf* regulatory regions (Figure 7-6). Using this system I aim to assess the development or abortive development of the haematopoietic system through both visual assessment and genome-wide profiling.

Figure 7-6 Schematic of proposed generation of *scl* knockout embryos, using genome editing techniques to replace the *scl* coding region with specific reporter gene cassettes.



In conclusion I have profiled and visualised a key developmental cell type, with great academic and clinical value, from a true *in vivo* vertebrate model. I have generated a transgenic zebrafish line that has permitted me to carry out transcriptional and open chromatin profiling at early stages of development with spatial specificity, which is not possible using culture systems. Comparison of these profiles has revealed key functional differences within the early *scl*<sup>+</sup> population, with sufficient detail to contribute to context specific regulatory network models. High resolution *in vivo* imaging has enabled me to relate my profiling results back into the embryo and visualise the dynamics and development of the haematopoietic and vascular system. Single cell sequencing of cellular RNA has identified subpopulations with differing profiles co-existing in the 10ss anterior *scl* population.

The tools generated in this project enable the *in vivo* study of a key cell type over a range of developmental stages and conditions. I am keen to continue working on this project, using ChIP and lineage tracing techniques to further develop our knowledge of these key progenitor populations. A greater understanding of the developmental decisions under the control of this master regulator will not only guide our understanding of leukemogenic transformation in many T-ALL cases but also potentially contribute to the development of regenerative techniques for the formation of haematopoietic and vascular engraftments.

## 8. Glossary

<b>2A</b>	self cleaving peptide sequence
<b>AGM</b>	Aorta-Gonad-Mesonephros
<b>ALM</b>	Anterior Lateral plate Mesoderm
<b>APLM</b>	Anterior of the Posterior Lateral plate Mesoderm
<b>ATAC</b>	Assay for Transposase Accessable Chromatin
<b>BAC</b>	Bacterial Artificial Chromosome
<b>bHLH</b>	basic Helix-Loop-Helix motif
<b>ChIP</b>	Chromatin ImmunoPrecipitation
<b>CHT</b>	Caudal Haematoipoietic Tissue
<b>CRISPR</b>	Clustered Regularly Interspaced Short Palindromic Repeats
<b>DA</b>	Dorsal Aorta
<b>DAPI</b>	4',6-diamidino-2-phenylindole
<b>dpf</b>	days post fertilisation
<b>EHT</b>	Endothelial to Haematopoietic Transition
<b>ERCC</b>	External RNA Control Consortium synthetic spike in standards
<b>FACS</b>	Fluorescence Activated Cell Sorting
<b>FLAG</b>	Polypeptide fusion protein tag
<b>FPKM</b>	Fragments Per Kilobase Mapped
<b>GAPDH</b>	Glyceraldehyde 3-phosphate dehydrogenase
<b>GATA</b>	GATA nucleotide sequence binding proteins
<b>GFP</b>	Green Fluorescent Protein
<b>GO</b>	Gene Ontology
<b>GTP</b>	Guanosine TriPhosphate
<b>HE</b>	Hemogenic Endothelium
<b>hpf</b>	hours post fertilisation
<b>HRP</b>	Horseradish Peroxidase
<b>HSC</b>	Haematopoietic Stem Cell
<b>ICC</b>	Immuno CytoChemistry
<b>ICM</b>	Intermediate Cell Mass
<b>L.A.-GFP</b>	Live Actin- GFP
<b>LPM</b>	Lateral Plate Mesoderm
<b>MAPK</b>	Mitogen Activated Protein Kinase
<b>MEL</b>	Murine ErythroLeukaemia
<b>NHEJ</b>	Non-Homologous End Joining
<b>PANTHER</b>	Protein Anylsis Through Evolutionary Relationships
<b>PBI</b>	Posterior Blood Island
<b>PKA</b>	Protein Kinase A
<b>PLM</b>	Posterior Lateral plate Mesoderm
<b>RT</b>	Room Temperature
<b>ss</b>	somite stage
<b>T-ALL</b>	T-cell Acute Lymphoblastic Leukaemia
<b>TCR</b>	T-Cell Receptor
<b>UTR</b>	UnTranslated Region
<b>VEGF</b>	Vascular Endothelial Growth Factor
<b>YFP</b>	Yellow Fluorescent Protein

## 9. List of Figures

Figure 1-1- Schematic of mesoderm formation, by Wolpert and Tickle, 2010 <sup>6</sup> .....	1-1
Figure 1-2- Schematic of vertebrate vasculogenesis, modified from Risau and Flamme, 1995 <sup>22</sup> .....	1-3
Figure 1-3- Vertebrate VEGFR classes, adapted from Bussmann <i>et al.</i> 2007 <sup>47</sup> .....	1-6
Figure 1-4- Model of Scl protein complex interactions with DNA in different cellular contexts, by El Omari <i>et al.</i> 2013 <sup>120</sup> .....	1-17
Figure 1-5- Synteny of the Scl locus across five vertebral species, by Gottgens <i>et al.</i> 2002 <sup>155</sup> .....	1-20
Figure 1-6 Schematic of haematopoietic development in the zebrafish highlighting the contexts of investigation in this project.....	1-33
Figure 2-1 PCR screening for F <sub>0</sub> s with germline transmission of the <i>scl-citrine</i> transgene.....	2-41
Figure 3-1 Schematic of the zc104B7 locus and the intended structure of a C-terminally tagged <i>scl</i> transgene	3-1
Figure 3-2 Endogenous <i>scl</i> expression patterns as determined by <i>in situ</i> hybridization.....	3-5
Figure 3-3 Initial results from injection of <i>sclBAC:scl-flag-tev-avi-2A-citrine-sv40pA</i> plus <i>tol2</i> mRNA into F <sub>0</sub> generation embryos.....	3-6
Figure 3-4 Fluorescent screening for <i>sclBAC:scl-flag-tev-avi-2A-citrine-sv40pA</i> transgenic F <sub>0</sub> s with germline transmission of transgene.....	3-7
Figure 3-5 Electrophoretic gel image of RT-PCR products, probing for transgenic transcripts.....	3-9
Figure 3-6 Time series showing <i>citrine</i> expression in the posterior of the 10-16ss <i>sclBAC:scl-flag-tev-avi-2A-citrine-sv40pA</i> transgenic F <sub>1</sub> embryo.....	3-10
Figure 3-7 Time series showing <i>citrine</i> expression in the posterior of a 16-20ss <i>sclBAC:scl-flag-tev-avi-2A-citrine-sv40pA</i> F <sub>1</sub> embryo.....	3-11
Figure 3-8 Time series showing <i>scl</i> driven <i>citrine</i> expression in the anterior of a 14-22ss <i>sclBAC:scl-flag-tev-avi-2A-citrine-sv40pA</i> transgenic F <sub>1</sub> embryo.....	3-12
Figure 3-9 Preliminary FACS isolation of <i>citrine</i> <sup>+</sup> populations.....	3-14
Figure 3-10 Schematic detailing the collection of spatially separated <i>scl</i> -expressing populations from zebrafish embryos.....	3-16
Figure 4-1 Scatter plots of replicates and PCA analysis of early <i>scl</i> expressing cells.....	4-20
Figure 4-2 Bar chart showing the overall number of genes expressed in each <i>scl</i> <sup>+</sup> context.....	4-21
Figure 4-3 Volcano plot showing the distribution of differentially expressed genes in 10ss <i>scl</i> <sup>+</sup> cells.....	4-24
Figure 4-4 Flat mounted 10ss embryos following <i>in situ</i> hybridization for a range of differentially expressed genes identified from RNA-seq datasets produced in this project.....	4-25
Figure 4-5 GO enrichment analysis of genes enriched in the 10ss anterior <i>scl</i> <sup>+</sup> population compared to the 10ss posterior <i>scl</i> <sup>+</sup> population.....	4-27
Figure 4-6 Biological and protein class analysis of genes enriched in the 10ss posterior <i>scl</i> expressing population.....	4-29

Figure 4-7 Biological gene ontology and protein class analysis of genes expressed at the 10ss in both the anterior and posterior <i>scf</i> expressing populations.....	4-32
Figure 4-8 Distribution of differentially expressed genes in <i>scf</i> <sup>+</sup> posterior cells.....	4-33
Figure 4-9 Biological gene ontology analysis genes enriched in the 20ss posterior <i>scf</i> <sup>+</sup> population.....	4-34
Figure 4-10 Biological gene ontology analysis genes enriched in the 20ss posterior <i>scf</i> <sup>+</sup> population. ....	4-36
Figure 4-11 Protein class and gene ontology analysis of genes commonly expressed in the 10 and 20ss posterior <i>scf</i> <sup>+</sup> populations.....	4-39
Figure 4-12 Venn diagram showing the distribution of genes expressed in all four early <i>scf</i> <sup>+</sup> contexts.....	4-40
Figure 4-13 Biological gene ontology analysis of genes expressed by all four <i>scf</i> <sup>+</sup> populations studied.....	4-41
Figure 4-14 Scatter plot of genes with the same expression level across all four early <i>scf</i> <sup>+</sup> contexts.....	4-43
Figure 4-15 Distribution of differentially expressed genes over the four early <i>scf</i> <sup>+</sup> contexts.....	4-47
Figure 4-16 Over-represented tissue-specific gene ontology terms for genes enriched in the 10ss anterior <i>scf</i> expressing population.....	4-49
Figure 4-17 Rho signalling enriched in the 10ss anterior <i>scf</i> expressing population.....	4-50
Figure 4-18 Over-represented tissue-specific gene ontology terms for genes enriched in the 10ss posterior <i>scf</i> expressing population.....	4-60
Figure 4-19 Over-represented tissue-specific gene ontology terms for genes enriched in the 20ss anterior <i>scf</i> expressing population.....	4-70
Figure 4-20 Over-represented tissue-specific gene ontology terms for genes enriched in the 20ss posterior <i>scf</i> expressing population.....	4-83
Figure 5-1 Overview of ATAC sequencing and mapping for early <i>scf</i> <sup>+</sup> cell populations .....	5-1
Figure 5-2 Distribution and overlap of non-promoter ATAC peaks in temporal comparisons of <i>scf</i> <sup>+</sup> cellular contexts.....	5-4
Figure 5-3 Distribution and overlap of non-promoter ATAC peaks in spatial comparisons of <i>scf</i> <sup>+</sup> cellular contexts.....	5-6
Figure 5-4 Venn diagram of context-specific binding site motif enrichment in non-promoter peaks .....	5-11
Figure 5-5 Open chromatin maps at the <i>gata1</i> locus of each of the contexts studied.....	5-12
Figure 6-1 Schematic of potential cellular diversity that could produce varied biological GO profiles. ....	6-16
Figure 6-2 Expression levels of <i>gata5</i> in early <i>scf</i> <sup>+</sup> populations. ....	6-18
Figure 6-3 Hyperspectral imaging of the 20ss embryo with <i>gata5</i> <sup>+</sup> and <i>scf</i> <sup>+</sup> populations labelled.....	6-20
Figure 6-4 Hyperspectral imaging of a double transgenic 10ss embryo produced from a cross of the Tg( <i>gata5</i> BAC: <i>lacZ</i> -GFP) and the Tg( <i>scf</i> BAC: <i>scf</i> -flag- <i>tev</i> - <i>ani</i> -2A- <i>citrine</i> - <i>sv40pA</i> ) transgenic lines.....	6-22
Figure 6-5 Expression of <i>kdrl</i> in early <i>scf</i> expressing populations .....	6-23
Figure 6-6 Double transgenic embryo produced from a cross of Tg( <i>kdrl</i> BAC: <i>gfp</i> ) and Tg( <i>scf</i> BAC: <i>scf</i> -flag- <i>tev</i> - <i>ani</i> -2A- <i>citrine</i> - <i>sv40pA</i> ) transgenic lines, fluorescently marking the <i>kdrl</i> and <i>scf</i> expressing populations at the 20ss. ....	6-24
Figure 6-7 Magnified image of double transgenic embryo, illustrating the variation in <i>kdrl</i> and <i>scf</i> expression patterns at the 20ss.....	6-25

Figure 6-8 Expression of fluorescent reporters for <i>sc/</i> and <i>kdr/</i> populations at the 10ss.....	6-27
Figure 6-9 Variation of <i>kdr/</i> expression patterns in the 10ss anterior.....	6-28
Figure 6-10 Single cell RNA sequencing read data per sample.....	6-29
Figure 6-11 PCA analysis of single cell transcriptome data, before and after quality control and filtering....	6-30
Figure 6-12 Modelling technical error and standard biological fluctuation, to assess the extent of transcriptome variability within the 10ss anterior <i>sc/</i> <sup>+</sup> population.....	6-32
Figure 6-13 Normalisation of variance and relating to biological GO function.....	6-33
Figure 6-14 Background distribution of the first principal component variance magnitude.....	6-34
Figure 6-15 PAGODA analysis of single cell 10ss anterior <i>sc/</i> <sup>+</sup> RNA-seq.....	6-36
Figure 6-16 Heat maps of expression level distribution of the top <u>expressed</u> genes of the 10ss anterior <i>sc/</i> expressing population.....	6-38
Figure 6-17 Annotated heat map of expression level distribution of the top 150 <u>expressed</u> genes of the 10ss anterior <i>sc/</i> expressing population.....	6-39
Figure 7-1 Time-lapse images of the beating heart from a 30hpf <i>sc/</i> -transgenic embryo .....	7-46
Figure 7-2 Schematic of BirA mediated biotin chromatin precipitation.....	7-48
Figure 7-3 Summary of top enriched gene ontology terms for each <i>sc/</i> <sup>+</sup> population investigated in this project .....	7-49
Figure 7-4 ATAC and RNA seq reads mapped to the zebrafish <i>sc/</i> loci, for each of the contexts studied in this project.....	7-51
Figure 7-5 Proposed cellular activity of early <i>sc/</i> <sup>+</sup> progenitor cells .....	7-54
Figure 7-6 Schematic of proposed generation of <i>sc/</i> knockout embryos, using genome editing techniques to replace the <i>sc/</i> coding region with specific reporter gene cassettes.....	7-57

## 10. List of Tables

Table 2-1 Results of single cell quality control. Samples to be omitted. ....	2-51
Table 2-2 Table of primers .....	2-53
Table 4-1 Erythroid factors enriched in the 10ss posterior <i>scf</i> expressing population. ....	4-30
Table 4-2 Genes associated with the GO term oxygen transport enriched in the 20ss posterior <i>scf</i> <sup>+</sup> population .....	4-37
Table 4-3 Genes associated with a molecular GO term of transcription factor and a biological GO term of haematopoiesis expressed by all four <i>scf</i> <sup>+</sup> populations studied. ....	4-42
Table 4-4 Haematopoietic genes showing common expression levels across the four early <i>scf</i> expressing populations studied. ....	4-44
Table 4-5 Embryonic haematopoietic genes enriched in the 10ss anterior <i>scf</i> expressing population. ....	4-51
Table 4-6 Genes annotated with the GO term myeloid cell differentiation enriched in the 10ss anterior <i>scf</i> expressing population. ....	4-53
Table 4-7 Genes annotated with the GO term angiogenesis enriched in the 10ss anterior <i>scf</i> expressing population.....	4-55
Table 4-8 Top enriched genes in the 10ss anterior <i>scf</i> expressing population. ....	4-57
Table 4-9 Genes annotated with the GO term oxygen transport enriched in the 10ss posterior <i>scf</i> expressing population.....	4-61
Table 4-10 Genes annotated with the GO term erythroid development enriched in the 10ss posterior <i>scf</i> expressing population. ....	4-62
Table 4-11 Genes annotated as involved in blood vessel morphogenesis enriched in the 10ss posterior <i>scf</i> expressing population. ....	4-64
Table 4-12 Genes annotated as histone lysine-methyl-transferases enriched in the 10ss posterior <i>scf</i> expressing population.....	4-66
Table 4-13 Top enriched genes in the 10ss posterior <i>scf</i> expressing population.....	4-69
Table 4-14 Genes annotated with the GO term oxygen transport enriched in the 20ss anterior <i>scf</i> expressing population.....	4-72
Table 4-15 Genes annotated with the GO term erythroid differentiationenriched in the 20ss anterior <i>scf</i> expressing population. ....	4-73
Table 4-16 Genes annotated with the GO term embryonic haematopoiesis enriched in the 20ss anterior <i>scf</i> expressing population. ....	4-74
Table 4-17 Genes annotated with the GO term myeloid differentiation enriched in the 20ss anterior <i>scf</i> expressing population. ....	4-77
Table 4-18 Genes annotated with the GO term endothelial cell migration enriched in the 20ss anterior <i>scf</i> expressing population. ....	4-78
Table 4-19 Genes annotated with the GO term angiogenic factors enriched in the 20ss anterior <i>scf</i> expressing population.....	4-80

Table 4-20 Top enriched genes in the 20ss anterior <i>scI</i> expressing population.....	4-82
Table 4-21 Genes annotated with the GO term oxygen transporter enriched in the 20ss posterior <i>scI</i> expressing population.....	4-85
Table 4-22 Genes annotated with the GO term embryonic haematopoiesis enriched in the 20ss posterior <i>scI</i> expressing population.....	4-86
Table 4-23 Genes annotated with the GO term erythroid differentiation enriched in the 20ss posterior <i>scI</i> expressing population.....	4-87
Table 4-24 Genes annotated with the GO term blood vessel morphogenesis enriched in the 20ss posterior <i>scI</i> expressing population.....	4-88
Table 4-25 Histone methyl transferase genes enriched in the 20ss posterior <i>scI</i> expressing population.....	4-90
Table 4-26 Top enriched genes in the 20ss posterior <i>scI</i> expressing population.....	4-92
Table 5-1 Distribution of promoter associated ATAC-seq peaks, between the four contexts studied. ....	5-2
Table 5-2 Distribution of <u>non-promoter</u> associated ATAC-seq peaks, between the four <i>scI</i> <sup>+</sup> contexts studied. ....	5-3
Table 6-1 Top 50 most over-dispersed genes in single cell RNA-seq datasets.....	6-42
Table 6-2 Variance of gene lists produced through GO analysis or from <i>de novo</i> evaluation of over-dispersion. ....	6-43

## 11. Bibliography

- 1 Ebendal, T. Migratory mesoblast cells in the young chick embryo examined by scanning electron microscopy. *Zoon* **4** (1976).
- 2 Wakely, J. & England, M. A. Scanning electron microscopy (SEM) of the chick embryo primitive streak. *Differentiation; research in biological diversity* **7**, 181-186 (1977).
- 3 Flamme, I. Edge cell migration in the extraembryonic mesoderm of the chick embryo. An experimental and morphological study. *Anatomy and embryology* **176**, 477-491 (1987).
- 4 Mayer, B. W., Jr. & Packard, D. S., Jr. A study of the expansion of the chick area vasculosa. *Dev Biol* **63**, 335-351 (1978).
- 5 Gonzalez-Crussi, F. Vasculogenesis in the chick embryo. An ultrastructural study. *American Journal of Anatomy* **130**, 441-459, doi:10.1002/aja.1001300406 (1971).
- 6 Wolpert, L. a. T., C. Principles of Development. *Oxford University Press* (2010).
- 7 W., H. Lecithoblast und Angioblast der Werbeltiere. *Abhandl. Math-Phys. Ges. Wiss.* **26**, 171-328 (1900).
- 8 Murray, P. D. F. The Development in vitro of the Blood of the Early Chick Embryo. *Proceedings of the Royal Society of London B: Biological Sciences* **111**, 497-521, doi:10.1098/rspb.1932.0070 (1932).
- 9 Kinder, S. J. *et al.* The orderly allocation of mesodermal cells to the extraembryonic structures and the anteroposterior axis during gastrulation of the mouse embryo. *Development* **126**, 4691-4701 (1999).
- 10 Choi, K., Kennedy, M., Kazarov, A., Papadimitriou, J. C. & Keller, G. A common precursor for hematopoietic and endothelial cells. *Development* **125**, 725-732 (1998).
- 11 Huber, T. L., Kouskoff, V., Fehling, H. J., Palis, J. & Keller, G. Haemangioblast commitment is initiated in the primitive streak of the mouse embryo. *Nature* **432**, 625-630, doi:10.1038/nature03122 (2004).
- 12 Lancrin, C. *et al.* The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature* **457**, 892-895, doi:10.1038/nature07679 (2009).
- 13 Mitrani, E., Gruenbaum, Y., Shohat, H. & Ziv, T. Fibroblast growth factor during mesoderm induction in the early chick embryo. *Development* **109**, 387-393 (1990).
- 14 Furuta, C. *et al.* Discordant developmental waves of angioblasts and hemangioblasts in the early gastrulating mouse embryo. *Development* **133**, 2771-2779, doi:10.1242/dev.02440 (2006).
- 15 Xiong, J.-W. Molecular and Developmental Biology of the Hemangioblast. *Developmental dynamics : an official publication of the American Association of Anatomists* **237**, 1218-1231, doi:10.1002/dvdy.21542 (2008).
- 16 Padrón-Barthe, L. *et al.* Clonal analysis identifies hemogenic endothelium as the source of the blood-endothelial common lineage in the mouse embryo. *Blood* **124**, 2523-2532, doi:10.1182/blood-2013-12-545939 (2014).
- 17 Gering, M., Rodaway, A. R. F., Göttgens, B., Patient, R. K. & Green, A. R. The SCL gene specifies haemangioblast development from early mesoderm. *EMBO Journal* **17**, 4029-4045 (1998).
- 18 Gering, M., Yamada, Y., Rabbitts, T. H. & Patient, R. K. Lmo2 and Scl/Tal1 convert non-axial mesoderm into haemangioblasts which differentiate into endothelial cells in the absence of Gata1. *Development* **130**, 6187-6199, doi:10.1242/dev.00875 (2003).

- 19 Chung, Y. S. *et al.* Lineage analysis of the hemangioblast as defined by FLK1 and SCL expression. *Development* **129**, 5511-5520, doi:10.1242/dev.00149 (2002).
- 20 Liu, F. & Patient, R. Genome-Wide Analysis of the Zebrafish ETS Family Identifies Three Genes Required for Hemangioblast Differentiation or Angiogenesis. *Circulation Research* **103**, 1147-1154, doi:10.1161/circresaha.108.179713 (2008).
- 21 Vogeli, K. M., Jin, S. W., Martin, G. R. & Stainier, D. Y. R. A common progenitor for haematopoietic and endothelial lineages in the zebrafish gastrula. *Nature* **443**, 337-339, doi:10.1038/nature05045 (2006).
- 22 Risau, W. & Flamme, I. Vasculogenesis. *Annual Review of Cell and Developmental Biology* **11**, 73-91, doi:doi:10.1146/annurev.cb.11.110195.000445 (1995).
- 23 Medvinsky, A., Rybtsov, S. & Taoudi, S. Embryonic origin of the adult hematopoietic system: Advances and questions. *Development* **138**, 1017-1031 (2011).
- 24 Coffin, J. D. & Poole, T. J. Embryonic vascular development: immunohistochemical identification of the origin and subsequent morphogenesis of the major vessel primordia in quail embryos. *Development* **102**, 735-748 (1988).
- 25 Sabin, F. R. Studies on the origin of blood vessels and of red corpuscles as seen in the living blastoderm of the chick during the second day of incubation. *Contributions to Embryology* **9**, 213-262 (1920).
- 26 Folkman, J. & Shing, Y. Angiogenesis. *The Journal of biological chemistry* **267**, 10931-10934 (1992).
- 27 Keller, G. Clonal analysis of hematopoietic stem cell development in vivo. *Current topics in microbiology and immunology* **177**, 41-57 (1992).
- 28 Osawa, M., Hanada, K., Hamada, H. & Nakauchi, H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* **273**, 242-245 (1996).
- 29 Flamme, I., von Reutern, M., Drexler, H. C., Syed-Ali, S. & Risau, W. Overexpression of vascular endothelial growth factor in the avian embryo induces hypervascularization and increased vascular permeability without alterations of embryonic pattern formation. *Dev Biol* **171**, 399-414, doi:10.1006/dbio.1995.1291 (1995).
- 30 Yamaguchi, T. P., Dumont, D. J., Conlon, R. A., Breitman, M. L. & Rossant, J. flk-1, an flt-related receptor tyrosine kinase is an early marker for endothelial cell precursors. *Development* **118**, 489-498 (1993).
- 31 Millauer, B. *et al.* High affinity VEGF binding and developmental expression suggest Flk-1 as a major regulator of vasculogenesis and angiogenesis. *Cell* **72**, 835-846 (1993).
- 32 Eichmann, A., Marcelle, C., Breant, C. & Le Douarin, N. M. Two molecules related to the VEGF receptor are expressed in early endothelial cells during avian embryonic development. *Mechanisms of development* **42**, 33-48 (1993).
- 33 Busmann, J., Bakkers, J. & Schulte-Merker, S. Early Endocardial Morphogenesis Requires Scl/Tal1. *PLoS Genet* **3**, e140, doi:10.1371/journal.pgen.0030140 (2007).
- 34 Neufeld, G., Cohen, T., Gengrinovitch, S. & Poltorak, Z. Vascular endothelial growth factor (VEGF) and its receptors. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **13**, 9-22 (1999).
- 35 Gerber, H. P. *et al.* VEGF is required for growth and survival in neonatal mice. *Development* **126**, 1149-1159 (1999).
- 36 Lawson, N. D., Vogel, A. M. & Weinstein, B. M. sonic hedgehog and vascular endothelial growth factor act upstream of the Notch pathway during arterial endothelial differentiation. *Dev Cell* **3**, 127-136 (2002).

- 37 Senger, D. *et al.* Tumor cells secrete a vascular permeability factor that promotes accumulation of ascites fluid. *Science* **219**, 983-985, doi:10.1126/science.6823562 (1983).
- 38 Hobson, B. & Denekamp, J. Endothelial proliferation in tumours and normal tissues: continuous labelling studies. *British Journal of Cancer* **49**, 405-413 (1984).
- 39 Engerman, R. L., Pfaffenbach, D. & Davis, M. D. Cell turnover of capillaries. *Laboratory investigation; a journal of technical methods and pathology* **17**, 738-743 (1967).
- 40 Pham, V. N. *et al.* Combinatorial function of ETS transcription factors in the developing vasculature. *Developmental Biology* **303**, 772-783 (2007).
- 41 Sumanas, S. *et al.* Interplay among Etsrp/ER71, Scl, and Alk8 signaling controls endothelial and myeloid cell formation. *Blood* **111**, 4500-4510, doi:10.1182/blood-2007-09-110569 (2008).
- 42 Sumanas, S. & Lin, S. Ets1-Related Protein Is a Key Regulator of Vasculogenesis in Zebrafish. *PLoS Biol* **4**, e10, doi:10.1371/journal.pbio.0040010 (2005).
- 43 Sharrocks, A. D., Brown, A. L., Ling, Y. & Yates, P. R. The ETS-domain transcription factor family. *Int J Biochem Cell Biol* **29**, 1371-1387 (1997).
- 44 Thompson, M. A. *et al.* The cloche and spadetail Genes Differentially Affect Hematopoiesis and Vasculogenesis. *Developmental Biology* **197**, 248-269 (1998).
- 45 Vlaeminck-Guillem, V. *et al.* The Ets family member Erg gene is expressed in mesodermal tissues and neural crests at fundamental steps during mouse embryogenesis. *Mechanisms of development* **91**, 331-335 (2000).
- 46 De Val, S. *et al.* Combinatorial Regulation of Endothelial Gene Expression by Ets and Forkhead Transcription Factors. *Cell* **135**, 1053-1064, doi:10.1016/j.cell.2008.10.049.
- 47 Bussmann, J., Bakkers, J. & Schulte-Merker, S. Early endocardial morphogenesis requires Scl/Tal1. *PLoS Genetics* **3**, 1425-1437, doi:10.1371/journal.pbio.0030314; (2007).
- 48 Lampugnani, M. G. *et al.* A novel endothelial-specific membrane protein is a marker of cell-cell contacts. *J Cell Biol* **118**, 1511-1522 (1992).
- 49 Baldwin, H. S. *et al.* Platelet endothelial cell adhesion molecule-1 (PECAM-1/CD31): alternatively spliced, functionally distinct isoforms expressed during mammalian cardiovascular development. *Development* **120**, 2539-2553 (1994).
- 50 Fina, L. *et al.* Expression of the CD34 gene in vascular endothelial cells. *Blood* **75**, 2417-2426 (1990).
- 51 Young, P. E., Baumhueter, S. & Lasky, L. A. The sialomucin CD34 is expressed on hematopoietic cells and blood vessels during murine development. *Blood* **85**, 96-105 (1995).
- 52 Abraham, S. *et al.* VE-Cadherin-Mediated Cell-Cell Interaction Suppresses Sprouting via Signaling to MLC2 Phosphorylation. *Current Biology* **19**, 668-674 (2009).
- 53 Bentley, K. *et al.* The role of differential VE-cadherin dynamics in cell rearrangement during angiogenesis. *Nat Cell Biol* **16**, 309-321, doi:10.1038/ncb2926 (2014).
- 54 Corada, M. *et al.* Vascular endothelial-cadherin is an important determinant of microvascular integrity in vivo. *Proc Natl Acad Sci U S A* **96**, 9815-9820 (1999).
- 55 Mitchell, I. C., Brown, T. S., Terada, L. S., Amatruda, J. F. & Nwariaku, F. E. Effect of Vascular Cadherin Knockdown on Zebrafish Vasculature during Development. *PLoS ONE* **5**, e8807, doi:10.1371/journal.pone.0008807 (2010).
- 56 Valge-Archer, V. E. *et al.* The LIM protein RBTN2 and the basic helix-loop-helix protein TAL1 are present in a complex in erythroid cells. *Proceedings of the National Academy of Sciences* **91**, 8617-8621 (1994).

- 57 Sadler, I., Crawford, A. W., Michelsen, J. W. & Beckerle, M. C. Zyxin and cCRP: two interactive LIM domain proteins associated with the cytoskeleton. *J Cell Biol* **119**, 1573-1587 (1992).
- 58 Schmeichel, K. L. & Beckerle, M. C. The LIM domain is a modular protein-binding interface. *Cell* **79**, 211-219 (1994).
- 59 Lecuyer, E. *et al.* Protein stability and transcription factor complex assembly determined by the SCL-LMO2 interaction. *The Journal of biological chemistry* **282**, 33649-33658, doi:10.1074/jbc.M703939200 (2007).
- 60 Boehm, T., Foroni, L., Kaneko, Y., Perutz, M. F. & Rabbitts, T. H. The rhombotin family of cysteine-rich LIM-domain oncogenes: distinct members are involved in T-cell translocations to human chromosomes 11p15 and 11p13. *Proc Natl Acad Sci U S A* **88**, 4367-4371 (1991).
- 61 Royer-Pokora, B., Loos, U. & Ludwig, W. D. TTG-2, a new gene encoding a cysteine-rich protein with the LIM motif, is overexpressed in acute T-cell leukaemia with the t(11;14)(p13;q11). *Oncogene* **6**, 1887-1893 (1991).
- 62 Foroni, L. *et al.* The rhombotin gene family encode related LIM-domain proteins whose differing expression suggests multiple roles in mouse development. *Journal of molecular biology* **226**, 747-761 (1992).
- 63 Warren, A. J. *et al.* The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development. *Cell* **78**, 45-57 (1994).
- 64 Begley, C. G. *et al.* The gene SCL is expressed during early hematopoiesis and encodes a differentiation-related DNA-binding motif. *Proc Natl Acad Sci U S A* **86**, 10128-10132 (1989).
- 65 Kurtzberg, J., Bigner, S. H. & Hershfield, M. S. Establishment of the DU.528 human lymphohemopoietic stem cell line. *J Exp Med* **162**, 1561-1578 (1985).
- 66 Hershfield, M. S. *et al.* Conversion of a stem cell leukemia from a T-lymphoid to a myeloid phenotype induced by the adenosine deaminase inhibitor 2'-deoxycoformycin. *Proc Natl Acad Sci U S A* **81**, 253-257 (1984).
- 67 Begley, C. G. *et al.* Chromosomal translocation in a human leukemic stem-cell line disrupts the T-cell antigen receptor delta-chain diversity region and results in a previously unreported fusion transcript. *Proc Natl Acad Sci U S A* **86**, 2031-2035 (1989).
- 68 Aplan, P. D. *et al.* The SCL gene is formed from a transcriptionally complex locus. *Mol Cell Biol* **10**, 6426-6435 (1990).
- 69 Begley, C. G. *et al.* Molecular Cloning and Chromosomal Localization of the Murine Homolog of the Human Helix-Loop-Helix Gene SCL. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 869-873 (1991).
- 70 Begley, C. G. *et al.* Structure of the gene encoding the murine SCL protein. *Gene* **138**, 93-99 (1994).
- 71 Breit, T. M., Wolvers-Tettero, I. L. & van Dongen, J. J. Lineage specific demethylation of tal-1 gene breakpoint region determines the frequency of tal-1 deletions in alpha beta lineage T-cells. *Oncogene* **9**, 1847-1853 (1994).
- 72 Visvader, J., Begley, C. G. & Adams, J. M. Differential expression of the LYL, SCL and E2A helix-loop-helix genes within the hemopoietic system. *Oncogene* **6**, 187-194 (1991).
- 73 Green, A. R., Lints, T., Visvader, J., Harvey, R. & Begley, C. G. SCL is coexpressed with GATA-1 in hemopoietic cells but is also expressed in developing brain. *Oncogene* **7**, 653-660 (1992).
- 74 Mouthon, M. A. *et al.* Expression of tal-1 and GATA-binding proteins during human hematopoiesis. *Blood* **81**, 647-655 (1993).

- 75 Brown, L. *et al.* Site-specific recombination of the tal-1 gene is a common  
occurrence in human T cell leukemia. *Embo j* **9**, 3343-3351 (1990).
- 76 Collazo-Garcia, N., Scherer, P. & Aplan, P. D. Cloning and characterization of a  
murine SIL gene. *Genomics* **30**, 506-513, doi:10.1006/geno.1995.1271 (1995).
- 77 Aplan, P. D. *et al.* Disruption of the human SCL locus by "illegitimate" V-(D)-J  
recombinase activity. *Science* **250**, 1426-1429 (1990).
- 78 Macintyre, E. A., Smit, L., Ritz, J., Kirsch, I. R. & Strominger, J. L. Disruption of  
the SCL locus in T-lymphoid malignancies correlates with commitment to the T-  
cell receptor alpha beta lineage. *Blood* **80**, 1511-1520 (1992).
- 79 Finger, L. R. *et al.* Involvement of the TCL5 gene on human chromosome 1 in T-  
cell leukemia and melanoma. *Proc Natl Acad Sci U S A* **86**, 5039-5043 (1989).
- 80 Aplan, P. D., Nakahara, K., Orkin, S. H. & Kirsch, I. R. The SCL gene product: a  
positive regulator of erythroid differentiation. *The EMBO Journal* **11**, 4073-4081  
(1992).
- 81 Lyons, S. E. *et al.* A nonsense mutation in zebrafish gata1 causes the bloodless  
phenotype in vlad tepes. *Proceedings of the National Academy of Sciences* **99**, 5454-  
5459, doi:10.1073/pnas.082695299 (2002).
- 82 Belele, C. L. *et al.* Differential requirement for Gata1 DNA binding and  
transactivation between primitive and definitive stages of hematopoiesis in  
zebrafish. *Blood* **114**, 5162-5172, doi:10.1182/blood-2009-05-224709 (2009).
- 83 Pevny, L. *et al.* Development of hematopoietic cells lacking transcription factor  
GATA-1. *Development* **121**, 163-172 (1995).
- 84 Martin, D. I. & Orkin, S. H. Transcriptional activation and DNA binding by the  
erythroid factor GF-1/NF-E1/Eryf 1. *Genes Dev* **4**, 1886-1898 (1990).
- 85 Juarez, M. A., Su, F., Chun, S., Kiel, M. J. & Lyons, S. E. Distinct roles for SCL  
in erythroid specification and maturation in zebrafish. *Journal of Biological Chemistry*  
**280**, 41636-41644, doi:10.1074/jbc.M507998200 (2005).
- 86 Shivdasanl, R. A., Mayer, E. L. & Orkin, S. H. Absence of blood formation in  
mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature* **373**, 432-434  
(1995).
- 87 Robb, L. *et al.* Absence of yolk sac hematopoiesis from mice with a targeted  
disruption of the scl gene. *Proc Natl Acad Sci U S A* **92**, 7075-7079 (1995).
- 88 Weiss, M. J., Keller, G. & Orkin, S. H. Novel insights into erythroid development  
revealed through in vitro differentiation of GATA-1 embryonic stem cells. *Genes  
& Development* **8**, 1184-1197, doi:10.1101/gad.8.10.1184 (1994).
- 89 Pevny, L. *et al.* Erythroid differentiation in chimaeric mice blocked by a targeted  
mutation in the gene for transcription factor GATA-1. *Nature* **349**, 257-260,  
doi:10.1038/349257a0 (1991).
- 90 Porcher, C. *et al.* The T cell leukemia oncoprotein SCL/tal-1 is essential for  
development of all hematopoietic lineages. *Cell* **86**, 47-57 (1996).
- 91 Robb, L. & Begley, C. G. The helix-loop-helix gene SCL: Implicated in T-cell  
acute lymphoblastic leukaemia and in normal haematopoietic development.  
*International Journal of Biochemistry and Cell Biology* **28**, 609-618, doi:10.1016/1357-  
2725(96)00006-4 (1996).
- 92 Herblot, S., Steff, A. M., Hugo, P., Aplan, P. D. & Hoang, T. SCL and LMO1  
alter thymocyte differentiation: inhibition of E2A-HEB function and pre-T alpha  
chain expression. *Nat Immunol* **1**, 138-144, doi:10.1038/77819 (2000).
- 93 Visvader, J. E., Fujiwara, Y. & Orkin, S. H. Unsuspected role for the T-cell  
leukemia protein SCL/tal-1 in vascular development. *Genes and Development* **12**,  
473-479 (1998).

- 94 Sanchez, M. *et al.* An SCL 3' enhancer targets developing endothelium together with embryonic and adult haematopoietic progenitors. *Development* **126**, 3891-3904 (1999).
- 95 Sanchez, M. J., Bockamp, E. O., Miller, J., Gambardella, L. & Green, A. R. Selective rescue of early haematopoietic progenitors in Scl(-/-) mice by expressing Scl under the control of a stem cell enhancer. *Development* **128**, 4815-4827 (2001).
- 96 Drake, C. J., Brandt, S. J., Trusk, T. C. & Little, C. D. TAL1/SCL is expressed in endothelial progenitor cells/angioblasts and defines a dorsal-to-ventral gradient of vasculogenesis. *Dev Biol* **192**, 17-30, doi:10.1006/dbio.1997.8751 (1997).
- 97 Patterson, L. J., Gering, M. & Patient, R. *Scl is required for dorsal aorta as well as blood formation in zebrafish embryos*. Vol. 105 (2005).
- 98 Stainier, D. Y., Weinstein, B. M., Detrich, H. W., 3rd, Zon, L. I. & Fishman, M. C. Cloche, an early acting zebrafish gene, is required by both the endothelial and hematopoietic lineages. *Development* **121**, 3141-3150 (1995).
- 99 Liao, E. C. *et al.* SCL/Tal-1 transcription factor acts downstream of cloche to specify hematopoietic and vascular progenitors in zebrafish. *Genes and Development* **12**, 621-626 (1998).
- 100 Drake, C. J. & Fleming, P. A. Vasculogenesis in the day 6.5 to 9.5 mouse embryo. *Blood* **95**, 1671-1679 (2000).
- 101 Lawson, N. D. & Weinstein, B. M. In vivo imaging of embryonic vascular development using transgenic zebrafish. *Dev Biol* **248**, 307-318 (2002).
- 102 Van Handel, B. *et al.* Scl represses cardiomyogenesis in prospective hemogenic endothelium and endocardium. *Cell* **150**, 590-605, doi:10.1016/j.cell.2012.06.026 (2012).
- 103 Pulford, K. *et al.* Expression of TAL-1 proteins in human tissues. *Blood* **85**, 675-684 (1995).
- 104 Sinclair, A. M. *et al.* Distinct 5' SCL Enhancers Direct Transcription to Developing Brain, Spinal Cord, and Endothelium: Neural Expression Is Mediated by GATA Factor Binding Sites. *Developmental Biology* **209**, 128-142 (1999).
- 105 Jin, H. *et al.* The 5' zebrafish scl promoter targets transcription to the brain, spinal cord, and hematopoietic and endothelial progenitors. *Dev Dyn* **235**, 60-67, doi:10.1002/dvdy.20613 (2006).
- 106 Ogilvy, S. *et al.* The SCL +40 enhancer targets the midbrain together with primitive and definitive hematopoiesis and is regulated by SCL and GATA proteins. *Molecular and Cellular Biology* **27**, 7206-7219, doi:10.1128/mcb.00931-07 (2007).
- 107 van Eekelen, J. A. *et al.* Expression pattern of the stem cell leukaemia gene in the CNS of the embryonic and adult mouse. *Neuroscience* **122**, 421-436 (2003).
- 108 Bockamp, E. O. *et al.* Distinct mechanisms direct SCL/tal-1 expression in erythroid cells and CD34 positive primitive myeloid cells. *The Journal of biological chemistry* **272**, 8781-8790 (1997).
- 109 Bockamp, E. O. *et al.* Transcriptional regulation of the stem cell leukemia gene by PU.1 and Elf-1. *Journal of Biological Chemistry* **273**, 29032-29042, doi:10.1074/jbc.273.44.29032 (1998).
- 110 George, K. M. *et al.* Embryonic expression and cloning of the murine GATA-3 gene. *Development* **120**, 2673-2686 (1994).
- 111 Engel, J. D. *et al.* cis and trans regulation of tissue-specific transcription. *Journal of cell science. Supplement* **16**, 21-31 (1992).

- 112 Courtial, N. *et al.* Tal1 regulates osteoclast differentiation through suppression of  
the master regulator of cell fusion DC-STAMP. *FASEB Journal* **26**, 523-532,  
doi:10.1096/fj.11-190850 (2012).
- 113 Yamane, T. *et al.* Sequential requirements for SCL/tal-1, GATA-2, macrophage  
colony-stimulating factor, and osteoclast differentiation factor/osteoprotegerin  
ligand in osteoclast development. *Exp Hematol* **28**, 833-840 (2000).
- 114 Murre, C., McCaw, P. S. & Baltimore, D. A new DNA binding and dimerization  
motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc  
proteins. *Cell* **56**, 777-783 (1989).
- 115 Mellentin, J. D., Smith, S. D. & Cleary, M. L. lyl-1, a novel gene altered by  
chromosomal translocation in T cell leukemia, codes for a protein with a helix-  
loop-helix DNA binding motif. *Cell* **58**, 77-83 (1989).
- 116 Porcher, C., Liao, E. C., Fujiwara, Y., Zon, L. I. & Orkin, S. H. Specification of  
hematopoietic and vascular development by the bHLH transcription factor SCL  
without direct DNA binding. *Development* **126**, 4603-4615 (1999).
- 117 Hsu, H. L., Wadman, I., Tsan, J. T. & Baer, R. Positive and negative  
transcriptional control by the TAL1 helix-loop-helix protein. *Proc Natl Acad Sci U  
S A* **91**, 5947-5951 (1994).
- 118 Hsu, H. L., Cheng, J. T., Chen, Q. & Baer, R. Enhancer-binding activity of the  
tal-1 oncoprotein in association with the E47/E12 helix-loop-helix proteins. *Mol  
Cell Biol* **11**, 3037-3042 (1991).
- 119 Massari, M. E. & Murre, C. Helix-loop-helix proteins: regulators of transcription  
in eucaryotic organisms. *Mol Cell Biol* **20**, 429-440 (2000).
- 120 El Omari, K. *et al.* Structural Basis for LMO2-Driven Recruitment of the  
SCL:E47bHLH Heterodimer to Hematopoietic-Specific Transcriptional Targets.  
*Cell Reports* **4**, 135-147 (2013).
- 121 Ferrando, A. A. *et al.* Gene expression signatures define novel oncogenic  
pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* **1**, 75-87 (2002).
- 122 Ferrando, A. A. *et al.* Biallelic transcriptional activation of oncogenic transcription  
factors in T-cell acute lymphoblastic leukemia. *Blood* **103**, 1909-1911,  
doi:10.1182/blood-2003-07-2577 (2004).
- 123 Larson, R. C. *et al.* Protein dimerization between Lmo2 (Rbtn2) and Tal1 alters  
thymocyte development and potentiates T cell tumorigenesis in transgenic mice.  
*The EMBO Journal* **15**, 1021-1027 (1996).
- 124 Mead, P. E., Deconinck, A. E., Huber, T. L., Orkin, S. H. & Zon, L. I. Primitive  
erythropoiesis in the Xenopus embryo: the synergistic role of LMO-2, SCL and  
GATA-binding proteins. *Development* **128**, 2301-2308 (2001).
- 125 Wadman, I. A. *et al.* The LIM - only protein Lmo2 is a bridging molecule  
assembling an erythroid, DNA - binding complex which includes the TAL1,  
E47, GATA - 1 and Ldb1/NLI proteins. *The EMBO Journal* **16**, 3145-3157,  
doi:10.1093/emboj/16.11.3145 (1997).
- 126 Kassouf, M. T. *et al.* Genome-wide identification of TAL1's functional targets:  
Insights into its mechanisms of action in primary erythroid cells. *Genome Research*  
**20**, 1064-1083 (2010).
- 127 Osada, H., Grutz, G., Axelson, H., Forster, A. & Rabbitts, T. H. Association of  
erythroid transcription factors: complexes involving the LIM protein RBTN2 and  
the zinc-finger protein GATA1. *Proceedings of the National Academy of Sciences of the  
United States of America* **92**, 9585-9589 (1995).
- 128 Fujiwara, Y., Browne, C. P., Cunniff, K., Goff, S. C. & Orkin, S. H. Arrested  
development of embryonic red cell precursors in mouse embryos lacking  
transcription factor GATA-1. *Proc Natl Acad Sci U S A* **93**, 12355-12358 (1996).

- 129 Lécuyer, E. *et al.* The SCL complex regulates c-kit expression in hematopoietic cells through functional interaction with Sp1. *Blood* **100**, 2430-2440, doi:10.1182/blood-2002-02-0568 (2002).
- 130 Ono, Y., Fukuhara, N. & Yoshie, O. TAL1 and LIM-only proteins synergistically induce retinaldehyde dehydrogenase 2 expression in T-cell acute lymphoblastic leukemia by acting as cofactors for GATA3. *Molecular and Cellular Biology* **18**, 6939-6950 (1998).
- 131 Cohen-Kaminsky, S. *et al.* Chromatin immunoselection defines a TAL-1 target gene. *EMBO Journal* **17**, 5151-5160, doi:10.1093/emboj/17.17.5151 (1998).
- 132 Schlaeger, T. M. *et al.* Decoding hematopoietic specificity in the helix-loop-helix domain of the transcription factor SCL/Tal-1. *Molecular and Cellular Biology* **24**, 7491-7502, doi:10.1128/mcb.24.17.7491-7502.2004 (2004).
- 133 Kassouf, M. T., Chagraoui, H., Vyas, P. & Porcher, C. Differential use of SCL/TAL-1 DNA-binding domain in developmental hematopoiesis. *Blood* **112**, 1056-1067, doi:10.1182/blood-2007-12-128900 (2008).
- 134 K M Draheim<sup>1</sup>, N. H., Y Yang<sup>1</sup>, E Arous<sup>1</sup>, J Calvo<sup>2</sup> and M A Kelliher<sup>1</sup>. A DNA-binding mutant of TAL1 cooperates with LMO2 to cause T cell leukemia in mice. *Oncogene* **30**, 1252–1260 (2011).
- 135 Schuh, A. H. *et al.* ETO-2 associates with SCL in erythroid cells and megakaryocytes and provides repressor functions in erythropoiesis. *Molecular and Cellular Biology* **25**, 10235-10250, doi:10.1128/mcb.25.23.10235-10250.2005 (2005).
- 136 Palomero, T. *et al.* Transcriptional regulatory networks downstream of TAL1/SCL in T-cell acute lymphoblastic leukemia. *Blood* **108**, 986-992, doi:10.1182/blood-2005-08-3482 (2006).
- 137 Wilson, N. K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532-544, doi:10.1016/j.stem.2010.07.016 (2010).
- 138 Pali, C. G. *et al.* Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *EMBO Journal* **30**, 494-509, doi:10.1038/emboj.2010.342 (2011).
- 139 Sanda, T. *et al.* Core Transcriptional Regulatory Circuit Controlled by the TAL1 Complex in Human T Cell Acute Lymphoblastic Leukemia. *Cancer Cell* **22**, 209-221 (2012).
- 140 Aplan, P. D. *et al.* An scl gene product lacking the transactivation domain induces bony abnormalities and cooperates with LMO1 to generate T-cell malignancies in transgenic mice. *EMBO Journal* **16**, 2408-2419, doi:10.1093/emboj/16.9.2408 (1997).
- 141 Schoenebeck, J. J., Keegan, B. R. & Yelon, D. Vessel and blood specification override cardiac potential in anterior mesoderm. *Dev Cell* **13**, 254-267, doi:10.1016/j.devcel.2007.05.012 (2007).
- 142 Simoes, F. C., Peterkin, T. & Patient, R. Fgf differentially controls cross-antagonism between cardiac and haemangioblast regulators. *Development* **138**, 3235-3245, doi:10.1242/dev.059634 (2011).
- 143 O'Neil, J., Shank, J., Cusson, N., Murre, C. & Kelliher, M. TAL1/SCL induces leukemia by inhibiting the transcriptional activity of E47/HEB. *Cancer Cell* **5**, 587-596, doi:10.1016/j.ccr.2004.05.023 (2004).
- 144 Kros, G. *et al.* Transcription factor SCL is required for c-kit expression and c-Kit function in hemopoietic cells. *Journal of Experimental Medicine* **188**, 439-450, doi:10.1084/jem.188.3.439 (1998).

- 145 Vitelli, L. *et al.* A pentamer transcriptional complex including tal-1 and  
retinoblastoma protein downmodulates c-kit expression in normal erythroblasts.  
*Mol Cell Biol* **20**, 5330-5342 (2000).
- 146 Huang, S., Qiu, Y., Stein, R. W. & Brandt, S. J. p300 functions as a transcriptional  
coactivator for the TAL1/SCL oncoprotein. *Oncogene* **18**, 4958-4967,  
doi:10.1038/sj.onc.1202889 (1999).
- 147 Kee, B. L., Arias, J. & Montminy, M. R. Adaptor-mediated recruitment of RNA  
polymerase II to a signal-dependent activator. *The Journal of biological chemistry* **271**,  
2373-2375 (1996).
- 148 Nakajima, T. *et al.* RNA Helicase A Mediates Association of CBP with RNA  
Polymerase II. *Cell* **90**, 1107-1112 (1997).
- 149 Abraham, S. E., Lobo, S., Yaciuk, P., Wang, H. G. & Moran, E. p300, and p300-  
associated proteins, are components of TATA-binding protein (TBP) complexes.  
*Oncogene* **8**, 1639-1647 (1993).
- 150 Yang, X. J., Ogryzko, V. V., Nishikawa, J., Howard, B. H. & Nakatani, Y. A  
p300/CBP-associated factor that competes with the adenoviral oncoprotein  
E1A. *Nature* **382**, 319-324, doi:10.1038/382319a0 (1996).
- 151 Huang, S. & Brandt, S. J. mSin3A regulates murine erythroleukemia cell  
differentiation through association with the TAL1 (or SCL) transcription factor.  
*Mol Cell Biol* **20**, 2248-2259 (2000).
- 152 Taunton, J., Hassig, C. A. & Schreiber, S. L. A mammalian histone deacetylase  
related to the yeast transcriptional regulator Rpd3p. *Science* **272**, 408-411 (1996).
- 153 Y Li, C. D., X Hu, B Patel, X Fu, Y Qiu, M Brand, K Zhao and S Huang.  
Dynamic interaction between TAL1 oncoprotein and LSD1 regulates TAL1  
function in hematopoiesis and leukemogenesis. *Oncogene* **31**, 5007–5018 (2012).
- 154 Hu, X. *et al.* LSD1-mediated epigenetic modification is required for TAL1  
function and hematopoiesis. *Proc Natl Acad Sci U S A* **106**, 10141-10146,  
doi:10.1073/pnas.0900437106 (2009).
- 155 Göttgens, B. *et al.* Transcriptional regulation of the stem cell leukemia gene (SCL)  
- Comparative analysis of five vertebrate SCL loci. *Genome Research* **12**, 749-759,  
doi:10.1101/gr.45502 (2002).
- 156 Elefanty, A. G., Begley, C. G., Hartley, L., Papaevangeliou, B. & Robb, L. SCL  
expression in the mouse embryo detected with a targeted lacZ reporter gene  
demonstrates its localization to hematopoietic, vascular, and neural tissues. *Blood*  
**94**, 3754-3763 (1999).
- 157 Thoms, J. A. *et al.* ERG promotes T-acute lymphoblastic leukemia and is  
transcriptionally regulated in leukemic cells by a stem cell enhancer. *Blood* **117**,  
7079-7089, doi:10.1182/blood-2010-12-317990 (2011).
- 158 O'Neil, J. *et al.* Alu elements mediate MYB gene tandem duplication in human T-  
ALL. *The Journal of Experimental Medicine* **204**, 3059-3066,  
doi:10.1084/jem.20071637 (2007).
- 159 Lahortiga, I. *et al.* Duplication of the MYB oncogene in T cell acute  
lymphoblastic leukemia. *Nature genetics* **39**, 593-595, doi:10.1038/ng2025 (2007).
- 160 Kusy, S. *et al.* NKX3.1 is a direct TAL1 target gene that mediates proliferation of  
TAL1-expressing human T cell acute lymphoblastic leukemia. *J Exp Med* **207**,  
2141-2156, doi:10.1084/jem.20100745 (2010).
- 161 Clappier, E. *et al.* The C-MYB locus is involved in chromosomal translocation  
and genomic duplications in human T-cell acute leukemia (T-ALL), the  
translocation defining a new T-ALL subtype in very young children. *Blood* **110**,  
1251-1261, doi:10.1182/blood-2006-12-064683 (2007).

- 162 Hosoya-Ohmura, S. *et al.* An NK and T cell enhancer lies 280 kilobase pairs 3' to the *gata3* structural gene. *Mol Cell Biol* **31**, 1894-1904, doi:10.1128/mcb.05065-11 (2011).
- 163 Nottingham, W. T. *et al.* Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. *Blood* **110**, 4188-4197, doi:10.1182/blood-2007-07-100883 (2007).
- 164 Landry, J. R. *et al.* Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors. *Blood* **113**, 5783-5792, doi:10.1182/blood-2008-11-187757 (2009).
- 165 Gottgens, B. *et al.* cis-Regulatory remodeling of the SCL locus during vertebrate evolution. *Mol Cell Biol* **30**, 5741-5751, doi:10.1128/mcb.00870-10 (2010).
- 166 Barton, L. M. *et al.* Regulation of the stem cell leukemia (SCL) gene: A tale of two fishes. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 6747-6752, doi:10.1073/pnas.101532998 (2001).
- 167 Smith, A. M. *et al.* A novel mode of enhancer evolution: the Tal1 stem cell enhancer recruited a MIR element to specifically boost its activity. *Genome Res* **18**, 1422-1432, doi:10.1101/gr.077008.108 (2008).
- 168 Dhami, P. *et al.* Genomic approaches uncover increasing complexities in the regulatory landscape at the human SCL (TAL1) locus. *PLoS One* **5**, e9059, doi:10.1371/journal.pone.0009059 (2010).
- 169 Peng, S. S., Chen, C. Y., Xu, N. & Shyu, A. B. RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein. *Embo j* **17**, 3461-3470, doi:10.1093/emboj/17.12.3461 (1998).
- 170 Afouda, A. B., Reynaud-Deonauth, S., Mohun, T. & Spohr, G. Localized Xid3 mRNA activation in *Xenopus* embryos by cytoplasmic polyadenylation. *Mechanisms of development* **88**, 15-31 (1999).
- 171 Correia, N. C. *et al.* Novel TAL1 targets beyond protein-coding genes: identification of TAL1-regulated microRNAs in T-cell acute lymphoblastic leukemia. *Leukemia* **27**, 1603-1606, doi:10.1038/leu.2013.63 (2013).
- 172 Mansour, M. R. *et al.* The TAL1 complex targets the FBXW7 tumor suppressor by activating miR-223 in human T cell acute lymphoblastic leukemia. *J Exp Med* **210**, 1545-1557, doi:10.1084/jem.20122516 (2013).
- 173 Correia, N. C. *et al.* *microRNAs regulate TAL1 expression in T-cell acute lymphoblastic leukemia.* (2016).
- 174 Palamarchuk, A. *et al.* Akt phosphorylates Tal1 oncoprotein and inhibits its repressor activity. *Cancer Research* **65**, 4515-4519, doi:10.1158/0008-5472.can-05-0751 (2005).
- 175 Nie, L., Wu, H. & Sun, X. H. Ubiquitination and degradation of Tal1/SCL are induced by Notch signaling and depend on Skp2 and CHIP. *Journal of Biological Chemistry* **283**, 684-692, doi:10.1074/jbc.M704981200 (2008).
- 176 Lawson, N. D. & Wolfe, S. A. Forward and reverse genetic approaches for the analysis of vertebrate development in the zebrafish. *Dev Cell* **21**, 48-64, doi:10.1016/j.devcel.2011.06.007 (2011).
- 177 Yu, Y. *et al.* A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. *Theoretical and Applied Genetics* **101**, 1093-1099, doi:10.1007/s001220051584 (2000).
- 178 Hruscha, A. *et al.* Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. *Development (Cambridge)* **140**, 4982-4987 (2013).
- 179 Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotech* **31**, 227-229, doi:10.1038/nbt.2501 (2013).

- 180 Irion, U., Krauss, J. & Nüsslein-Volhard, C. Precise and efficient genome editing  
in zebrafish using the CRISPR/Cas9 system. *Development* **141**, 4827-4830,  
doi:10.1242/dev.115584 (2014).
- 181 Corey, D. R. & Abrams, J. M. Morpholino antisense oligonucleotides: tools for  
investigating vertebrate development. *Genome Biology* **2**, reviews1015.1011-  
reviews1015.1013 (2001).
- 182 Bugel, S. M., Tanguay, R. L. & Planchart, A. Zebrafish: A marvel of high-  
throughput biology for 21 century toxicology. *Current environmental health reports* **1**,  
341-352, doi:10.1007/s40572-014-0029-5 (2014).
- 183 Rennekamp, A. J. & Peterson, R. T. 15 years of zebrafish chemical screening.  
*Current opinion in chemical biology* **24**, 58-70, doi:10.1016/j.cbpa.2014.10.025 (2015).
- 184 Huiting, L. N., Laroche, F. & Feng, H. The Zebrafish as a Tool to Cancer Drug  
Discovery. *Austin journal of pharmacology and therapeutics* **3**, 1069 (2015).
- 185 North, T. E. *et al.* Prostaglandin E2 regulates vertebrate haematopoietic stem cell  
homeostasis. *Nature* **447**, 1007-1011, doi:10.1038/nature05883 (2007).
- 186 Cutler, C. *et al.* Prostaglandin-modulated umbilical cord blood hematopoietic  
stem cell transplantation. *Blood* **122**, 3074-3081, doi:10.1182/blood-2013-05-  
503177 (2013).
- 187 Davidson, A. J. *et al.* cdx4 mutants fail to specify blood progenitors and can be  
rescued by multiple hox genes. *Nature* **425**, 300-306, doi:10.1038/nature01973  
(2003).
- 188 Detrich, H. W. *et al.* Intraembryonic hematopoietic cell migration during  
vertebrate development. *Proceedings of the National Academy of Sciences of the United  
States of America* **92**, 10713-10717 (1995).
- 189 Dooley, K. A., Davidson, A. J. & Zon, L. I. Zebrafish scl functions  
independently in hematopoietic and endothelial development. *Developmental Biology*  
**277**, 522-536, doi:10.1016/j.ydbio.2004.09.004 (2005).
- 190 Bertrand, J. Y. *et al.* Definitive hematopoiesis initiates through a committed  
erythromyeloid progenitor in the zebrafish embryo. *Development* **134**, 4147-4156,  
doi:10.1242/dev.012385 (2007).
- 191 Bennett, C. M. *et al.* Myelopoiesis in the zebrafish, *Danio rerio*. *Blood* **98**, 643-651  
(2001).
- 192 Bertrand, J. Y., Kim, A. D., Teng, S. & Traver, D. CD41+ cmyb+ precursors  
colonize the zebrafish pronephros by a novel migration route to initiate adult  
hematopoiesis. *Development* **135**, 1853-1862, doi:10.1242/dev.015297 (2008).
- 193 Murayama, E. *et al.* Tracing hematopoietic precursor migration to successive  
hematopoietic organs during zebrafish development. *Immunity* **25**, 963-975,  
doi:10.1016/j.immuni.2006.10.015 (2006).
- 194 North, T. *et al.* Cbfa2 is required for the formation of intra-aortic hematopoietic  
clusters. *Development* **126**, 2563-2575 (1999).
- 195 Zovein, A. C. *et al.* Fate tracing reveals the endothelial origin of hematopoietic  
stem cells. *Cell Stem Cell* **3**, 625-636, doi:10.1016/j.stem.2008.09.018 (2008).
- 196 Eilken, H. M., Nishikawa, S. & Schroeder, T. Continuous single-cell imaging of  
blood generation from haemogenic endothelium. *Nature* **457**, 896-900,  
doi:10.1038/nature07760 (2009).
- 197 Kissa, K. *et al.* Live imaging of emerging hematopoietic stem cells and early  
thymus colonization. *Blood* **111**, 1147-1156, doi:10.1182/blood-2007-07-099499  
(2008).
- 198 Zhang, X. Y. & Rodaway, A. R. SCL-GFP transgenic zebrafish: in vivo imaging  
of blood and endothelial development and identification of the initial site of

- definitive hematopoiesis. *Dev Biol* **307**, 179-194, doi:10.1016/j.ydbio.2007.04.002 (2007).
- 199 Jin, H. *et al.* Definitive hematopoietic stem/progenitor cells manifest distinct differentiation output in the zebrafish VDA and PBI. *Development (Cambridge, England)* **136**, 647-654, doi:10.1242/dev.029637 (2009).
- 200 Trede, N. S., Zapata, A. & Zon, L. I. Fishing for lymphoid genes. *Trends in immunology* **22**, 302-307 (2001).
- 201 Jagadeeswaran, P. & Sheehan, J. P. Analysis of blood coagulation in the zebrafish. *Blood cells, molecules & diseases* **25**, 239-249 (1999).
- 202 Patterson, L. J. *et al.* The transcription factors Scl and Lmo2 act together during development of the hemangioblast in zebrafish. *Blood* **109**, 2389-2398 (2007).
- 203 Gering, M., Rodaway, A. R., Gottgens, B., Patient, R. K. & Green, A. R. The SCL gene specifies haemangioblast development from early mesoderm. *The EMBO journal* **17**, 4029-4045, doi:10.1093/emboj/17.14.4029 (1998).
- 204 Vogeli, K. M., Jin, S. W., Martin, G. R. & Stainier, D. Y. A common progenitor for haematopoietic and endothelial lineages in the zebrafish gastrula. *Nature* **443**, 337-339, doi:10.1038/nature05045 (2006).
- 205 Rhodes, J. *et al.* Interplay of Pu.1 and Gata1 Determines Myelo-Erythroid Progenitor Cell Fate in Zebrafish. *Developmental Cell* **8**, 97-108, doi:10.1016/j.devcel.2004.11.014.
- 206 Herbomel, P., Thisse, B. & Thisse, C. Zebrafish early macrophages colonize cephalic mesenchyme and developing brain, retina, and epidermis through a M-CSF receptor-dependent invasive process. *Dev Biol* **238**, 274-288, doi:10.1006/dbio.2001.0393 (2001).
- 207 Lieschke, G. J., Oates, A. C., Crowhurst, M. O., Ward, A. C. & Layton, J. E. Morphologic and functional characterization of granulocytes and macrophages in embryonic and adult zebrafish. *Blood* **98**, 3087-3096 (2001).
- 208 Hume, D. A. Plenary Perspective: The complexity of constitutive and inducible gene expression in mononuclear phagocytes. *Journal of Leukocyte Biology* **92**, 433-444, doi:10.1189/jlb.0312166 (2012).
- 209 Kawane, K. *et al.* Requirement of DNase II for definitive erythropoiesis in the mouse fetal liver. *Science* **292**, 1546-1549, doi:10.1126/science.292.5521.1546 (2001).
- 210 Gordon, S. & Taylor, P. R. Monocyte and macrophage heterogeneity. *Nature reviews. Immunology* **5**, 953-964, doi:10.1038/nri1733 (2005).
- 211 Ginhoux, F. *et al.* Fate mapping analysis reveals that adult microglia derive from primitive macrophages. *Science* **330**, 841-845, doi:10.1126/science.1194637 (2010).
- 212 Hoeffel, G. *et al.* Adult Langerhans cells derive predominantly from embryonic fetal liver monocytes with a minor contribution of yolk sac-derived macrophages. *The Journal of Experimental Medicine* **209**, 1167-1181, doi:10.1084/jem.20120340 (2012).
- 213 Hume, D. A. Macrophages as APC and the dendritic cell myth. *J Immunol* **181**, 5829-5835 (2008).
- 214 Chitu, V. & Stanley, E. R. Colony-stimulating factor-1 in immunity and inflammation. *Current opinion in immunology* **18**, 39-48, doi:10.1016/j.coi.2005.11.006 (2006).
- 215 Satpathy, A. T., Wu, X., Albring, J. C. & Murphy, K. M. Re(de)fining the dendritic cell lineage. *Nature immunology* **13**, 1145-1154, doi:10.1038/ni.2467 (2012).
- 216 Schulz, C. *et al.* A lineage of myeloid cells independent of Myb and hematopoietic stem cells. *Science* **336**, 86-90, doi:10.1126/science.1219179 (2012).

- 217 Galloway, J. L., Wingert, R. A., Thisse, C., Thisse, B. & Zon, L. I. Loss of gata1  
but not gata2 converts erythropoiesis to myelopoiesis in zebrafish embryos. *Dev*  
*Cell* **8**, 109-116, doi:10.1016/j.devcel.2004.12.001 (2005).
- 218 Berman, J. N., Kanki, J. P. & Look, A. T. Zebrafish as a model for myelopoiesis  
during embryogenesis. *Exp Hematol* **33**, 997-1006,  
doi:10.1016/j.exphem.2005.06.010 (2005).
- 219 Bernard, O. *et al.* A third tal-1 promoter is specifically used in human T cell  
leukemias. *J Exp Med* **176**, 919-925 (1992).
- 220 Calkhoven, C. F. *et al.* Translational control of SCL-isoform expression in  
hematopoietic lineage choice. *Genes Dev* **17**, 959-964, doi:10.1101/gad.251903  
(2003).
- 221 Bockamp, E. O. *et al.* Lineage-restricted regulation of the murine SCL/TAL-1  
promoter. *Blood* **86**, 1502-1514 (1995).
- 222 Qian, F. *et al.* Distinct functions for different scl isoforms in zebrafish primitive  
and definitive hematopoiesis. *PLoS biology* **5**, doi:10.1371/journal.pbio.0050132  
(2007).
- 223 Zhen, F., Lan, Y., Yan, B., Zhang, W. & Wen, Z. Hemogenic endothelium  
specification and hematopoietic stem cell maintenance employ distinct Scl  
isoforms. *Development (Cambridge)* **140**, 3977-3985, doi:10.1242/dev.097071 (2013).
- 224 Chen, M. J., Yokomizo, T., Zeigler, B. M., Dzierzak, E. & Speck, N. A. Runx1 is  
required for the endothelial to haematopoietic cell transition but not thereafter.  
*Nature* **457**, 887-891 (2009).
- 225 Ren, X., Gomez, G. A., Zhang, B. & Lin, S. Scl isoforms act downstream of etsrp  
to specify angioblasts and definitive hematopoietic stem cells. *Blood* **115**, 5338-  
5346, doi:10.1182/blood-2009-09-244640 (2010).
- 226 Fong, T. A. *et al.* SU5416 is a potent and selective inhibitor of the vascular  
endothelial growth factor receptor (Flk-1/KDR) that inhibits tyrosine kinase  
catalysis, tumor vascularization, and growth of multiple tumor types. *Cancer Res*  
**59**, 99-106 (1999).
- 227 Westerfield, M. The Zebrafish Book; A Guide for the Laboratory Use of  
Zebrafish (*Brachydanio rerio*). *University of Oregon Press Eugene* (1993).
- 228 Andrews, S. FastQC A Quality Control tool for High Throughput Sequence  
Data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
- 229 Buffalo, V. Scythe - A Bayesian adapter trimmer (version 0.994 BETA).  
<https://github.com/vsbuffalo/scythe> (2014).
- 230 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford,*  
*England)*, doi:10.1093/bioinformatics/bts635 (2012).
- 231 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and  
quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628,  
doi:10.1038/nmeth.1226 (2008).
- 232 Anders, S. & Huber, W. Differential expression analysis for sequence count data.  
*Genome Biology* **11**, 1-12, doi:10.1186/gb-2010-11-10-r106 (2010).
- 233 Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An Abundance of  
Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence  
Data. *PLoS computational biology* **5**, e1000598, doi:10.1371/journal.pcbi.1000598  
(2009).
- 234 Langmead, B. Aligning short sequencing reads with Bowtie. *Current protocols in*  
*bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.] CHAPTER*, Unit-  
11.17, doi:10.1002/0471250953.bi1107s32 (2010).
- 235 Broad Institute. PicardTools (v.1.83) MarkDuplicate package.  
<http://broadinstitute.github.io/picard/> (2015).

- 236 Ramírez, F., DüNDAR, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, doi:10.1093/nar/gku365 (2014).
- 237 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
- 238 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
- 239 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Meth* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).
- 240 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
- 241 Trompouki, E. *et al.* Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell* **147**, 577-589, doi:10.1016/j.cell.2011.09.044 (2011).
- 242 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181, doi:10.1038/nprot.2014.006 (2014).
- 243 Guo, M., Wang, H., Potter, S. S., Whitsett, J. A. & Xu, Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS computational biology* **11**, e1004575, doi:10.1371/journal.pcbi.1004575 (2015).
- 244 Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protocols* **8**, 1765-1786, doi:10.1038/nprot.2013.099 (2013).
- 245 Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Meth* **11**, 163-166, doi:10.1038/nmeth.2772 (2014).
- 246 Cai, Y. *et al.* Eto2/MTG16 and MTGR1 are heteromeric corepressors of the TAL1/SCL transcription factor in murine erythroid progenitors. *Biochemical and Biophysical Research Communications* **390**, 295-301, doi:10.1016/j.bbrc.2009.09.111 (2009).
- 247 Lausen, J. *et al.* Targets of the Tal1 transcription factor in erythrocytes: E2 ubiquitin conjugase regulation by Tal1. *Journal of Biological Chemistry* **285**, 5338-5346, doi:10.1074/jbc.M109.030296 (2010).
- 248 Ryan, M. D., King, A. M. & Thomas, G. P. Cleavage of foot-and-mouth disease virus polyprotein is mediated by residues located within a 19 amino acid sequence. *The Journal of general virology* **72 ( Pt 11)**, 2727-2732, doi:10.1099/0022-1317-72-11-2727 (1991).
- 249 Szymczak, A. L. *et al.* Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector. *Nat Biotech* **22**, 589-594 (2004).
- 250 Donnelly, M. L. *et al.* Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: a putative ribosomal 'skip'. *The Journal of general virology* **82**, 1013-1025, doi:10.1099/0022-1317-82-5-1013 (2001).
- 251 Yu, D. *et al.* An efficient recombination system for chromosome engineering in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5978-5983, doi:10.1073/pnas.100127597 (2000).
- 252 Bussmann, J. & Schulte-Merker, S. Rapid BAC selection for tol2-mediated transgenesis in zebrafish. *Development* **138**, 4327-4332, doi:10.1242/dev.068080 (2011).

- 253 Suster, M. L., Sumiyama, K. & Kawakami, K. Transposon-mediated BAC transgenesis in zebrafish and mice. *BMC genomics* **10**, 477, doi:10.1186/1471-2164-10-477 (2009).
- 254 Kawakami, K. Tol2: a versatile gene transfer vector in vertebrates. *Genome Biology* **8**, S7-S7, doi:10.1186/gb-2007-8-s1-s7 (2007).
- 255 Suster, M. L., Kikuta, H., Urasaki, A., Asakawa, K. & Kawakami, K. Transgenesis in zebrafish with the tol2 transposon system. *Methods in molecular biology (Clifton, N.J.)* **561**, 41-63, doi:10.1007/978-1-60327-019-9\_3 (2009).
- 256 Kawakami, K., Koga, A., Hori, H. & Shima, A. Excision of the tol2 transposable element of the medaka fish, *Oryzias latipes*, in zebrafish, *Danio rerio*. *Gene* **225**, 17-22 (1998).
- 257 Urasaki, A., Morvan, G. & Kawakami, K. Functional Dissection of the Tol2 Transposable Element Identified the Minimal cis-Sequence and a Highly Repetitive Sequence in the Subterminal Region Essential for Transposition. *Genetics* **174**, 639-649, doi:10.1534/genetics.106.060244 (2006).
- 258 Ni, J. *et al.* Active recombinant Tol2 transposase for gene transfer and gene discovery applications. *Mobile DNA* **7**, 6, doi:10.1186/s13100-016-0062-z (2016).
- 259 Meir, Y.-J. J. *et al.* Genome-wide target profiling of piggyBac and Tol2in HEK 293: pros and cons for gene discovery and gene therapy. *BMC Biotechnology* **11**, 28, doi:10.1186/1472-6750-11-28 (2011).
- 260 Kondrychyn, I., Garcia-Lecea, M., Emelyanov, A., Parinov, S. & Korzh, V. Genome-wide analysis of Tol2 transposon reintegration in zebrafish. *BMC genomics* **10**, 418-418, doi:10.1186/1471-2164-10-418 (2009).
- 261 Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLoS computational biology* **5**, e1000598, doi:10.1371/journal.pcbi.1000598 (2009).
- 262 Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protocols* **8**, 1551-1566, doi:10.1038/nprot.2013.092 (2013).
- 263 Tarafdar, A., Dobbin, E., Corrigan, P., Freeburn, R. & Wheadon, H. Canonical Wnt Signaling Promotes Early Hematopoietic Progenitor Formation and Erythroid Specification during Embryonic Stem Cell Differentiation. *PLoS ONE* **8**, e81030, doi:10.1371/journal.pone.0081030 (2013).
- 264 Sturgeon, C. M., Ditadi, A., Awong, G., Kennedy, M. & Keller, G. Wnt signaling controls the specification of definitive and primitive hematopoiesis from human pluripotent stem cells. *Nat Biotech* **32**, 554-561, doi:10.1038/nbt.2915 (2014).
- 265 Paluru, P. *et al.* The Negative Impact of Wnt Signaling on Megakaryocyte and Primitive Erythroid Progenitors Derived From Human Embryonic Stem Cells. *Stem cell research* **12**, 441-451, doi:10.1016/j.scr.2013.12.003 (2014).
- 266 Park, S. T. & Sun, X. H. The Tal1 oncoprotein inhibits E47-mediated transcription. Mechanism of inhibition. *Journal of Biological Chemistry* **273**, 7030-7037, doi:10.1074/jbc.273.12.7030 (1998).
- 267 Gu, Y., Jasti, A. C., Jansen, M. & Siefring, J. E. RhoH, a hematopoietic-specific Rho GTPase, regulates proliferation, survival, migration, and engraftment of hematopoietic progenitor cells. *Blood* **105**, 1467-1475, doi:10.1182/blood-2004-04-1604 (2005).
- 268 Nayak, R. C., Chang, K.-H., Vaitinadin, N.-S. & Cancelas, J. A. Rho GTPases control specific cytoskeleton-dependent functions of hematopoietic stem cells. *Immunological reviews* **256**, 10.1111/imr.12119, doi:10.1111/imr.12119 (2013).

- 269 Gu, Y. *et al.* Hematopoietic cell regulation by Rac1 and Rac2 guanosine triphosphatases. *Science* **302**, 445-449, doi:10.1126/science.1088485 (2003).
- 270 Ghiaur, G. *et al.* Rac1 is essential for intraembryonic hematopoiesis and for the initial seeding of fetal liver with definitive hematopoietic progenitor cells. *Blood* **111**, 3313-3321, doi:10.1182/blood-2007-08-110114 (2008).
- 271 Yang, F. C. *et al.* Rac and Cdc42 GTPases control hematopoietic stem cell shape, adhesion, migration, and mobilization. *Proc Natl Acad Sci U S A* **98**, 5614-5618, doi:10.1073/pnas.101546898 (2001).
- 272 El Omari, K. *et al.* Structure of the leukemia oncogene LMO2: Implications for the assembly of a hematopoietic transcription factor complex. *Blood* **117**, 2146-2156, doi:10.1182/blood-2010-07-293357 (2011).
- 273 Hamlett, I. *et al.* Characterization of megakaryocyte GATA1-interacting proteins: the corepressor ETO2 and GATA1 interact to regulate terminal megakaryocyte maturation. *Blood* **112**, 2738-2749, doi:10.1182/blood-2008-03-146605 (2008).
- 274 Grimes, H. L., Chan, T. O., Zweidler-McKay, P. A., Tong, B. & Tsichlis, P. N. The Gfi-1 proto-oncoprotein contains a novel transcriptional repressor domain, SNAG, and inhibits G1 arrest induced by interleukin-2 withdrawal. *Mol Cell Biol* **16**, 6263-6272 (1996).
- 275 Wei, W. *et al.* Gfi1.1 regulates hematopoietic lineage differentiation during zebrafish embryogenesis. *Cell research* **18**, 677-685, doi:10.1038/cr.2008.60 (2008).
- 276 Ransom, D. G. *et al.* Characterization of zebrafish mutants with defects in embryonic hematopoiesis. *Development* **123**, 311-319 (1996).
- 277 Liao, E. C. *et al.* Hereditary spherocytosis in zebrafish riesling illustrates evolution of erythroid beta-spectrin structure, and function in red cell morphogenesis and membrane stability. *Development* **127**, 5123-5132 (2000).
- 278 Lieschke, G. J. *et al.* Zebrafish SPI-1 (PU.1) Marks a Site of Myeloid Development Independent of Primitive Erythropoiesis: Implications for Axial Patterning. *Developmental Biology* **246**, 274-295 (2002).
- 279 Liang, D., Jia, W., Li, J., Li, K. & Zhao, Q. Retinoic Acid Signaling Plays a Restrictive Role in Zebrafish Primitive Myelopoiesis. *PLoS ONE* **7**, e30865, doi:10.1371/journal.pone.0030865 (2012).
- 280 Liongue, C., Hall, C. J., O'Connell, B. A., Crosier, P. & Ward, A. C. Zebrafish granulocyte colony-stimulating factor receptor signaling promotes myelopoiesis and myeloid cell migration. *Blood* **113**, 2535-2546, doi:10.1182/blood-2008-07-171967 (2009).
- 281 Klemsz, M. J., McKercher, S. R., Celada, A., Van Beveren, C. & Maki, R. A. The macrophage and B cell-specific transcription factor PU.1 is related to the *ets* oncogene. *Cell* **61**, 113-124, doi:10.1016/0092-8674(90)90219-5.
- 282 McKercher, S. R. *et al.* Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *The EMBO Journal* **15**, 5647-5658 (1996).
- 283 Scott, E., Simon, M., Anastasi, J. & Singh, H. Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science* **265**, 1573-1577, doi:10.1126/science.8079170 (1994).
- 284 Peterkin, T., Gibson, A. & Patient, R. Common genetic control of haemangioblast and cardiac development in zebrafish. *Development* **136**, 1465-1474, doi:10.1242/dev.032748 (2009).
- 285 Holtzinger, A. & Evans, T. Gata5 and Gata6 are functionally redundant in zebrafish for specification of cardiomyocytes. *Dev Biol* **312**, 613-622, doi:10.1016/j.ydbio.2007.09.018 (2007).

- 286 Peterkin, T., Gibson, A. & Patient, R. Redundancy and evolution of GATA  
factor requirements in development of the myocardium. *Dev Biol* **311**, 623-635,  
doi:10.1016/j.ydbio.2007.08.018 (2007).
- 287 Reiter, J. F. *et al.* Gata5 is required for the development of the heart and  
endoderm in zebrafish. *Genes & Development* **13**, 2983-2995 (1999).
- 288 Le, A. T., Yelon, D. & Stainier, D. Y. R. Hand2 Regulates Epithelial Formation  
during Myocardial Differentiation. *Current Biology* **15**, 441-446 (2005).
- 289 Ohneda, K. & Yamamoto, M. Roles of hematopoietic transcription factors  
GATA-1 and GATA-2 in the development of red blood cell lineage. *Acta  
haematologica* **108**, 237-245, doi:65660 (2002).
- 290 Gao, X. *et al.* Gata2 cis-element is required for hematopoietic stem cell  
generation in the mammalian embryo. *The Journal of Experimental Medicine* **210**,  
2833-2842, doi:10.1084/jem.20130733 (2013).
- 291 Kissa, K. & Herbomel, P. Blood stem cells emerge from aortic endothelium by a  
novel type of cell transition. *Nature* **464**, 112-115 (2010).
- 292 Brownlie, A. *et al.* Positional cloning of the zebrafish sauternes gene: a model for  
congenital sideroblastic anaemia. *Nature genetics* **20**, 244-250, doi:10.1038/3049  
(1998).
- 293 Bottomley, S. S., May, B. K., Cox, T. C., Cotter, P. D. & Bishop, D. F. Molecular  
defects of erythroid 5-aminolevulinic synthase in X-linked sideroblastic anemia.  
*Journal of Bioenergetics and Biomembranes* **27**, 161-168, doi:10.1007/bf02110031.
- 294 Larson, J. D. *et al.* Expression of VE-cadherin in zebrafish embryos: A new tool  
to evaluate vascular development. *Developmental Dynamics* **231**, 204-213,  
doi:10.1002/dvdy.20102 (2004).
- 295 Breviario, F. *et al.* Functional properties of human vascular endothelial cadherin  
(7B4/cadherin-5), an endothelium-specific cadherin. *Arteriosclerosis, thrombosis, and  
vascular biology* **15**, 1229-1239 (1995).
- 296 Anderson, H. *et al.* Hematopoietic stem cells develop in the absence of  
endothelial cadherin 5 expression. *Blood* **126**, 2811-2820, doi:10.1182/blood-2015-  
07-659276 (2015).
- 297 Lancrin, C. *et al.* The haemangioblast generates haematopoietic cells through a  
haemogenic endothelium stage. *Nature* **457**, 892-895, doi:10.1038/nature07679  
(2009).
- 298 Chen, M. J. *et al.* Erythroid/myeloid progenitors and hematopoietic stem cells  
originate from distinct populations of endothelial cells. *Cell Stem Cell* **9**, 541-552,  
doi:10.1016/j.stem.2011.10.003 (2011).
- 299 Garnaas, M. K. *et al.* Syx, a RhoA Guanine Exchange Factor, Is Essential for  
Angiogenesis In Vivo. *Circulation Research* **103**, 710-716,  
doi:10.1161/circresaha.108.181388 (2008).
- 300 Ernkvist, M. *et al.* The Amot/Patj/Syx signaling complex spatially controls RhoA  
GTPase activity in migrating endothelial cells. *Blood* **113**, 244-253,  
doi:10.1182/blood-2008-04-153874 (2009).
- 301 Brown, L. A. *et al.* Insights into early vasculogenesis revealed by expression of the  
ETS-domain transcription factor Fli-1 in wild-type and mutant zebrafish  
embryos. *Mechanisms of development* **90**, 237-252 (2000).
- 302 Covassin, L. *et al.* Global analysis of hematopoietic and vascular endothelial gene  
expression by tissue specific microarray profiling in zebrafish. *Developmental Biology*  
**299**, 551-562 (2006).
- 303 Spyropoulos, D. D. *et al.* Hemorrhage, impaired hematopoiesis, and lethality in  
mouse embryos carrying a targeted disruption of the Fli1 transcription factor. *Mol  
Cell Biol* **20**, 5643-5652 (2000).

- 304 Birdsey, G. M. *et al.* Transcription factor Erg regulates angiogenesis and  
endothelial apoptosis through VE-cadherin. *Blood* **111**, 3498-3506,  
doi:10.1182/blood-2007-08-105346 (2008).
- 305 Bahary, N. *et al.* Duplicate VegfA genes and orthologues of the KDR receptor  
tyrosine kinase family mediate vascular development in the zebrafish. *Blood* **110**,  
3627-3636, doi:10.1182/blood-2006-04-016378 (2007).
- 306 Bridge, G. *et al.* The microRNA-30 family targets DLL4 to modulate endothelial  
cell behavior during angiogenesis. *Blood* **120**, 5063-5072, doi:10.1182/blood-2012-  
04-423004 (2012).
- 307 Krueger, J. *et al.* Flt1 acts as a negative regulator of tip cell formation and  
branching morphogenesis in the zebrafish embryo. *Development* **138**, 2111-2120,  
doi:10.1242/dev.063933 (2011).
- 308 Rottbauer, W. *et al.* VEGF-PLC $\gamma$ 1 pathway controls cardiac contractility in the  
embryonic heart. *Genes & Development* **19**, 1624-1634, doi:10.1101/gad.1319405  
(2005).
- 309 Wakayama, Y., Fukuhara, S., Ando, K., Matsuda, M. & Mochizuki, N. Cdc42  
Mediates Bmp-Induced Sprouting Angiogenesis through Fmnl3-Driven  
Assembly of Endothelial Filopodia in Zebrafish. *Developmental Cell* **32**, 109-122  
(2015).
- 310 Cermenati, S. *et al.* Sox18 and Sox7 play redundant roles in vascular development.  
*Blood* **111**, 2657-2666, doi:10.1182/blood-2007-07-100412 (2008).
- 311 Griffin, K. J. *et al.* A conserved role for H15-related T-box transcription factors  
in zebrafish and Drosophila heart formation. *Dev Biol* **218**, 235-247,  
doi:10.1006/dbio.1999.9571 (2000).
- 312 Szeto, D. P., Griffin, K. J. & Kimelman, D. HrT is required for cardiovascular  
development in zebrafish. *Development* **129**, 5093-5101 (2002).
- 313 Zhong, T. P., Rosenberg, M., Mohideen, M. A., Weinstein, B. & Fishman, M. C.  
gridlock, an HLH gene required for assembly of the aorta in zebrafish. *Science*  
**287**, 1820-1824 (2000).
- 314 Zhong, T. P., Childs, S., Leu, J. P. & Fishman, M. C. Gridlock signalling pathway  
fashions the first embryonic artery. *Nature* **414**, 216-220 (2001).
- 315 Higgs, D. R. *et al.* A review of the molecular genetics of the human alpha-globin  
gene cluster. *Blood* **73**, 1081-1104 (1989).
- 316 Shaw, G. C. *et al.* Mitoferrin is essential for erythroid iron assimilation. *Nature*  
**440**, 96-100 (2006).
- 317 Weinstein, B. M. *et al.* Hematopoietic mutations in the zebrafish. *Development* **123**,  
303-309 (1996).
- 318 Childs, S. *et al.* Zebrafish dracula encodes ferrochelatase and its mutation  
provides a model for erythropoietic protoporphyria. *Current biology : CB* **10**, 1001-  
1004 (2000).
- 319 Towatari, M. *et al.* Regulation of GATA-2 Phosphorylation by Mitogen-activated  
Protein Kinase and Interleukin-3. *Journal of Biological Chemistry* **270**, 4101-4107  
(1995).
- 320 Crossley, M. & Orkin, S. H. Phosphorylation of the erythroid transcription factor  
GATA-1. *Journal of Biological Chemistry* **269**, 16589-16596 (1994).
- 321 Katzav, S., Martin-Zanca, D. & Barbacid, M. vav, a novel human oncogene  
derived from a locus ubiquitously expressed in hematopoietic cells. *Embo j* **8**,  
2283-2290 (1989).
- 322 Wulf, G. M., Adra, C. N. & Lim, B. Inhibition of hematopoietic development  
from embryonic stem cells by antisense vav RNA. *Embo j* **12**, 5065-5074 (1993).

- 323 Zhang, X. Y. & Rodaway, A. R. F. SCL-GFP transgenic zebrafish: In vivo  
imaging of blood and endothelial development and identification of the initial site  
of definitive hematopoiesis. *Developmental Biology* **307**, 179-194,  
doi:10.1016/j.ydbio.2007.04.002 (2007).
- 324 Huminiecki, L. & Bicknell, R. In silico cloning of novel endothelial-specific  
genes. *Genome Res* **10**, 1796-1806 (2000).
- 325 Verma, A. *et al.* Endothelial cell-specific chemotaxis receptor (ecscr) promotes  
angioblast migration during vasculogenesis and enhances VEGF receptor  
sensitivity. *Blood* **115**, 4614-4622, doi:10.1182/blood-2009-10-248856 (2010).
- 326 Martyn, U. & Schulte-Merker, S. Zebrafish neuropilins are differentially  
expressed and interact with vascular endothelial growth factor during embryonic  
vascular development. *Developmental Dynamics* **231**, 33-42, doi:10.1002/dvdy.20048  
(2004).
- 327 Hu, G. *et al.* A novel endothelial-specific heat shock protein HspA12B is required  
in both zebrafish development and endothelial functions in vitro. *Journal of Cell  
Science* **119**, 4117-4126, doi:10.1242/jcs.03179 (2006).
- 328 Kouzarides, T. Histone methylation in transcriptional control. *Current Opinion in  
Genetics & Development* **12**, 198-209 (2002).
- 329 Peterson, C. L. & Laniel, M.-A. Histones and histone modifications. *Current  
Biology* **14**, R546-R551 (2004).
- 330 Van Laarhoven, P. M. *et al.* Kabuki syndrome genes KMT2D and KDM6A:  
functional analyses demonstrate critical roles in craniofacial, heart and brain  
development. *Human Molecular Genetics* **24**, 4443-4453, doi:10.1093/hmg/ddv180  
(2015).
- 331 Martin, C. & Zhang, Y. The diverse functions of histone lysine methylation. *Nat  
Rev Mol Cell Biol* **6**, 838-849 (2005).
- 332 Laurenson, P. & Rine, J. Silencers, silencing, and heritable transcriptional states.  
*Microbiological Reviews* **56**, 543-560 (1992).
- 333 van Leeuwen, F., Gafken, P. R. & Gottschling, D. E. Dot1p Modulates Silencing  
in Yeast by Methylation of the Nucleosome Core. *Cell* **109**, 745-756 (2002).
- 334 Ng, H. H., Ciccone, D. N., Morshead, K. B., Oettinger, M. A. & Struhl, K.  
Lysine-79 of histone H3 is hypomethylated at silenced loci in yeast and  
mammalian cells: A potential mechanism for position-effect variegation.  
*Proceedings of the National Academy of Sciences* **100**, 1820-1825,  
doi:10.1073/pnas.0437846100 (2003).
- 335 Xu, F. *et al.* N-CoR is required for patterning the anterior-posterior axis of  
zebrafish hindbrain by actively repressing retinoid signaling. *Mechanisms of  
development* **126**, 771-780 (2009).
- 336 Li, J., Li, K., Dong, X., Liang, D. & Zhao, Q. Ncor1 and Ncor2 play essential but  
distinct roles in zebrafish primitive myelopoiesis. *Developmental Dynamics* **243**,  
1544-1553, doi:10.1002/dvdy.24181 (2014).
- 337 Chen, L.-M. *et al.* Cloning and characterization of a zebrafish homologue of  
human AQP1: a bifunctional water and gas channel. *American Journal of Physiology -  
Regulatory, Integrative and Comparative Physiology* **299**, R1163-R1174,  
doi:10.1152/ajpregu.00319.2010 (2010).
- 338 Rehn, K., Wong, K. S., Balciunas, D. & Sumanas, S. Zebrafish enhancer trap line  
recapitulates embryonic aquaporin 1a expression pattern in vascular endothelial  
cells. *The International journal of developmental biology* **55**, 613-618,  
doi:10.1387/ijdb.103249kp (2011).
- 339 Wei, M. *et al.* The over-expression of aquaporin-1 alters erythroid gene  
expression in human erythroleukemia K562 cells. *Tumour biology : the journal of the*

- International Society for Oncodevelopmental Biology and Medicine* **36**, 291-302, doi:10.1007/s13277-014-2614-5 (2015).
- 340 Wei, M., Shi, R., Jiang, L., Wang, N. & Ma, W. [Role of aquaporin-1 gene in erythroid differentiation of erythroleukemia K562 cells induced by retinoic acid]. *Nan fang yi ke da xue xue bao = Journal of Southern Medical University* **32**, 1689-1694 (2012).
- 341 Gilmour, K. M., Thomas, K., Esbaugh, A. J. & Perry, S. F. Carbonic anhydrase expression and CO<sub>2</sub> excretion during early development in zebrafish *Danio rerio*. *Journal of Experimental Biology* **212**, 3837-3845, doi:10.1242/jeb.034116 (2009).
- 342 Qian, F. *et al.* Microarray analysis of zebrafish cloche mutant using amplified cDNA and identification of potential downstream target genes. *Developmental Dynamics* **233**, 1163-1172, doi:10.1002/dvdy.20444 (2005).
- 343 Paw, B. H. *et al.* Cell-specific mitotic defect and dyserythropoiesis associated with erythroid band 3 deficiency. *Nature genetics* **34**, 59-64, doi:10.1038/ng1137 (2003).
- 344 Traver, D. *et al.* Transplantation and in vivo imaging of multilineage engraftment in zebrafish bloodless mutants. *Nat Immunol* **4**, 1238-1246, doi:10.1038/ni1007 (2003).
- 345 Shafizadeh, E. *et al.* Characterization of zebrafish merlot/chablis as non-mammalian vertebrate models for severe congenital anemia due to protein 4.1 deficiency. *Development* **129**, 4359-4370 (2002).
- 346 Eckfeldt, C. E. *et al.* Functional Analysis of Human Hematopoietic Stem Cell Gene Expression Using Zebrafish. *PLoS Biol* **3**, e254, doi:10.1371/journal.pbio.0030254 (2005).
- 347 He, Q. *et al.* Inflammatory signaling regulates hematopoietic stem and progenitor cell emergence in vertebrates. *Blood* **125**, 1098-1106, doi:10.1182/blood-2014-09-601542 (2015).
- 348 Stachura, D. L. *et al.* The zebrafish granulocyte colony-stimulating factors (Gcsfs): 2 paralogous cytokines and their roles in hematopoietic development and maintenance. *Blood* **122**, 3918-3928, doi:10.1182/blood-2012-12-475392 (2013).
- 349 Radomska, H. S. *et al.* CCAAT/enhancer binding protein alpha is a regulatory switch sufficient for induction of granulocytic development from bipotential myeloid progenitors. *Mol Cell Biol* **18**, 4301-4314 (1998).
- 350 Hohaus, S. *et al.* PU.1 (Spi-1) and C/EBP alpha regulate expression of the granulocyte-macrophage colony-stimulating factor receptor alpha gene. *Molecular and Cellular Biology* **15**, 5830-5845, doi:10.1128/mcb.15.10.5830 (1995).
- 351 Scott, L. M., Civin, C. I., Rorth, P. & Friedman, A. D. A novel temporal expression pattern of three C/EBP family members in differentiating myelomonocytic cells. *Blood* **80**, 1725-1735 (1992).
- 352 Liu, T. X. *et al.* Dominant-interfering C/EBP $\alpha$  stimulates primitive erythropoiesis in zebrafish. *Experimental Hematology* **35**, 230-239 (2007).
- 353 Chen, J. N. *et al.* Mutations affecting the cardiovascular system and other internal organs in zebrafish. *Development* **123**, 293-302 (1996).
- 354 Le, A. T. & Stainier, D. Y. R. Fibronectin Regulates Epithelial Organization during Myocardial Migration in Zebrafish. *Developmental Cell* **6**, 371-382 (2004).
- 355 Chiu, C.-H., Chou, C.-W., Takada, S. & Liu, Y.-W. Development and Fibronectin Signaling Requirements of the Zebrafish Interrenal Vessel. *PLoS ONE* **7**, e43040, doi:10.1371/journal.pone.0043040 (2012).
- 356 Leslie, J. D. *et al.* Endothelial signalling by the Notch ligand Delta-like 4 restricts angiogenesis. *Development* **134**, 839-844, doi:10.1242/dev.003244 (2007).
- 357 Quillien, A. *et al.* Distinct Notch signaling outputs pattern the developing arterial system. *Development* **141**, 1544-1552, doi:10.1242/dev.099986 (2014).

- 358 Siekmann, A. F. & Lawson, N. D. Notch signalling limits angiogenic cell  
behaviour in developing zebrafish arteries. *Nature* **445**, 781-784,  
doi:10.1038/nature05577 (2007).
- 359 Yelon, D., Horne, S. A. & Stainier, D. Y. Restricted expression of cardiac myosin  
genes reveals regulated aspects of heart tube assembly in zebrafish. *Dev Biol* **214**,  
23-37, doi:10.1006/dbio.1999.9406 (1999).
- 360 Gulliver, G. Observations on the sizes and shapes of the red corpuscles of the  
blood of vertebrates, with drawings of them to uniform scale, and extended and  
revised tables of measurement. *Proc. Zool. Soc. Lond.* **474**. (1875).
- 361 Weijs, B. G. M. W. *et al.* E2F7 and E2F8 promote angiogenesis through  
transcriptional activation of VEGFA in cooperation with HIF1. *The EMBO  
Journal* **31**, 3871-3884, doi:10.1038/emboj.2012.231 (2012).
- 362 Herzog, W., Müller, K., Huisken, J. & Stainier, D. Y. R. Genetic Evidence for a  
Noncanonical Function of Seryl-tRNA Synthetase in Vascular Development.  
*Circulation Research* **104**, 1260-1266, doi:10.1161/circresaha.108.191718 (2009).
- 363 Lento, W., Congdon, K., Voermans, C., Kritzik, M. & Reya, T. Wnt Signaling in  
Normal and Malignant Hematopoiesis. *Cold Spring Harbor Perspectives in Biology* **5**,  
a008011, doi:10.1101/cshperspect.a008011 (2013).
- 364 Kumano, K. *et al.* Notch1 but not Notch2 is essential for generating  
hematopoietic stem cells from endothelial cells. *Immunity* **18**, 699-711 (2003).
- 365 Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-  
access database of transcription factor binding profiles. *Nucleic Acids Research* **44**,  
D110-D115, doi:10.1093/nar/gkv1176 (2016).
- 366 Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B.  
JASPAR: an open-access database for eukaryotic transcription factor binding  
profiles. *Nucleic Acids Research* **32**, D91-D94, doi:10.1093/nar/gkh012 (2004).
- 367 Sarkar, A. & Hochedlinger, K. The Sox Family of Transcription Factors:  
Versatile Regulators of Stem and Progenitor Cell Fate. *Cell Stem Cell* **12**, 15-30  
(2013).
- 368 Schepers, G. E., Teasdale, R. D. & Koopman, P. Twenty pairs of Sox: Extent,  
homology, and nomenclature of the mouse and human Sox transcription factor  
gene families. *Developmental Cell* **3**, 167-170, doi:10.1016/S1534-5807(02)00223-X  
(2002).
- 369 Liber, D. *et al.* Epigenetic priming of a Pre-B Cell-Specific enhancer through  
binding of Sox2 and Foxd3 at the ESC stage. *Cell Stem Cell* **7**, 114-126,  
doi:10.1016/j.stem.2010.05.020 (2010).
- 370 Kim, I., Saunders, T. L. & Morrison, S. J. Sox17 Dependence Distinguishes the  
Transcriptional Regulation of Fetal from Adult Hematopoietic Stem Cells. *Cell*  
**130**, 470-483, doi:10.1016/j.cell.2007.06.011 (2007).
- 371 He, S., Kim, I., Lim, M. S. & Morrison, S. J. Sox17 expression confers self-  
renewal potential and fetal stem cell characteristics upon adult hematopoietic  
progenitors. *Genes and Development* **25**, 1613-1627, doi:10.1101/gad.2052911  
(2011).
- 372 Wang, G. L., Jiang, B. H., Rue, E. A. & Semenza, G. L. Hypoxia-inducible factor  
1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O<sub>2</sub> tension.  
*Proceedings of the National Academy of Sciences of the United States of America* **92**, 5510-  
5514 (1995).
- 373 Krock, B. L., Skuli, N. & Simon, M. C. Hypoxia-Induced Angiogenesis: Good  
and Evil. *Genes & Cancer* **2**, 1117-1133, doi:10.1177/1947601911423654 (2011).

- 374 Adelman, D. M., Maltepe, E. & Simon, M. C. in *Oxygen Sensing: Molecule to Man* (eds Sukhamay Lahiri, Naduri R. Prabhakar, & Robert E. Forster) 275-284 (Springer US, 2002).
- 375 Hong, W. X. *et al.* The Role of Hypoxia-Inducible Factor in Wound Healing. *Advances in wound care* **3**, 390-399, doi:10.1089/wound.2013.0520 (2014).
- 376 Goode, Debbie K. *et al.* Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Developmental Cell* **36**, 572-587, doi:10.1016/j.devcel.2016.01.024.
- 377 Dong, J. *et al.* Elucidation of a universal size-control mechanism in Drosophila and mammals. *Cell* **130**, 1120-1133, doi:10.1016/j.cell.2007.07.019 (2007).
- 378 Milton, C. C. *et al.* The Hippo pathway regulates hematopoiesis in Drosophila melanogaster. *Current biology : CB* **24**, 2673-2680, doi:10.1016/j.cub.2014.10.031 (2014).
- 379 Ruiz i Altaba, A. Gli proteins encode context-dependent positive and negative functions: implications for development and disease. *Development* **126**, 3205-3216 (1999).
- 380 Merchant, A., Joseph, G., Wang, Q., Brennan, S. & Matsui, W. Gli1 regulates the proliferation and differentiation of HSCs and myeloid progenitors. *Blood* **115**, 2391-2396, doi:10.1182/blood-2009-09-241703 (2010).
- 381 St-Jacques, B., Hammerschmidt, M. & McMahon, A. P. Indian hedgehog signaling regulates proliferation and differentiation of chondrocytes and is essential for bone formation. *Genes Dev* **13**, 2072-2086 (1999).
- 382 Dyer, M. A., Farrington, S. M., Mohn, D., Munday, J. R. & Baron, M. H. Indian hedgehog activates hematopoiesis and vasculogenesis and can respecify prospective neurectodermal cell fate in the mouse embryo. *Development* **128**, 1717-1730 (2001).
- 383 Hu, M. *et al.* Multilineage gene expression precedes commitment in the hemopoietic system. *Genes & Development* **11**, 774-785, doi:10.1101/gad.11.6.774 (1997).
- 384 Mercer, E. M. *et al.* Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors. *Immunity* **35**, 413-425, doi:10.1016/j.immuni.2011.06.013 (2011).
- 385 Rodaway, A. *et al.* Induction of the mesendoderm in the zebrafish germ ring by yolk cell-derived TGF-beta family signals and discrimination of mesoderm and endoderm by FGF. *Development* **126**, 3067-3078 (1999).
- 386 Reiter, J. F., Kikuchi, Y. & Stainier, D. Y. Multiple roles for Gata5 in zebrafish endoderm formation. *Development* **128**, 125-135 (2001).
- 387 Riedl, J. *et al.* Lifeact: a versatile marker to visualize F-actin. *Nature methods* **5**, 605, doi:10.1038/nmeth.1220 (2008).
- 388 Lou, X., Deshwar, A. R., Crump, J. G. & Scott, I. C. Smarcd3b and Gata5 promote a cardiac progenitor fate in the zebrafish embryo. *Development* **138**, 3113-3123, doi:10.1242/dev.064279 (2011).
- 389 Liao, W. *et al.* The zebrafish gene cloche acts upstream of a flk-1 homologue to regulate endothelial cell differentiation. *Development* **124**, 381-389 (1997).
- 390 Covassin, L. D., Villefranc, J. A., Kacergis, M. C., Weinstein, B. M. & Lawson, N. D. Distinct genetic interactions between multiple Vegf receptors are required for development of different blood vessel types in zebrafish. *Proceedings of the National Academy of Sciences* **103**, 6554-6559, doi:10.1073/pnas.0506886103 (2006).
- 391 Lee, Y. M. *et al.* Vascular endothelial growth factor receptor signaling is required for cardiac valve formation in zebrafish. *Developmental Dynamics* **235**, 29-37, doi:10.1002/dvdy.20559 (2006).

- 392 Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-  
cell differential expression analysis. *Nat Meth* **11**, 740-742,  
doi:10.1038/nmeth.2967 (2014).
- 393 Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and  
gene set overdispersion analysis. *Nat Methods* **13**, 241-244,  
doi:10.1038/nmeth.3734 (2016).
- 394 Wong, K. S. *et al.* Hedgehog signaling is required for differentiation of  
endocardial progenitors in zebrafish. *Developmental Biology* **361**, 377-391 (2012).
- 395 Wilson, N. K. *et al.* The transcriptional program controlled by the stem cell  
leukemia gene *Scl/Tal1* during early embryonic hematopoietic development.  
*Blood* **113**, 5456-5465, doi:10.1182/blood-2009-01-200048 (2009).
- 396 Tijssen, Marloes R. *et al.* Genome-wide Analysis of Simultaneous GATA1/2,  
RUNX1, FLI1, and SCL Binding in Megakaryocytes Identifies Hematopoietic  
Regulators. *Developmental Cell* **20**, 597-609, doi:10.1016/j.devcel.2011.04.008  
(2011).
- 397 Swiers, G., Patient, R. & Loose, M. Genetic regulatory networks programming  
hematopoietic stem cells and erythroid lineage specification. *Developmental Biology*  
**294**, 525-540, doi:10.1016/j.ydbio.2006.02.051 (2006).
- 398 Loose, M., Swiers, G. & Patient, R. Transcriptional networks regulating  
hematopoietic cell fate decisions. *Current Opinion in Hematology* **14**, 307-314,  
doi:10.1097/MOH.0b013e3281900eee (2007).
- 399 Davidson, E. H. *et al.* A genomic regulatory network for development. *Science* **295**,  
1669-1678, doi:10.1126/science.1069883 (2002).
- 400 Cripps, R. M. & Olson, E. N. Control of cardiac development by an  
evolutionarily conserved transcriptional network. *Dev Biol* **246**, 14-28,  
doi:10.1006/dbio.2002.0666 (2002).
- 401 Howard, M. L. & Davidson, E. H. cis-Regulatory control circuits in  
development. *Developmental Biology* **271**, 109-118 (2004).
- 402 Loose, M. & Patient, R. A genetic regulatory network for *Xenopus* mesendoderm  
formation. *Developmental Biology* **271**, 467-478 (2004).
- 403 Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains  
uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* **38**,  
1348-1354 (2006).
- 404 Ema, M. *et al.* Combinatorial effects of Flk1 and Tal1 on vascular and  
hematopoietic development in the mouse. *Genes & Development* **17**, 380-393,  
doi:10.1101/gad.1049803 (2003).
- 405 Zape, J. P. & Zovein, A. C. Hemogenic endothelium: origins, regulation, and  
implications for vascular biology. *Seminars in cell & developmental biology* **22**, 1036-  
1047, doi:10.1016/j.semcdb.2011.10.003 (2011).
- 406 Huber, T. L., Kouskoff, V., Fehling, H. J., Palis, J. & Keller, G. Haemangioblast  
commitment is initiated in the primitive streak of the mouse embryo. *Nature* **432**,  
625-630, doi:10.1038/nature03122 (2004).
- 407 Butko, E. *et al.* Gata2b is a restricted early regulator of hemogenic endothelium in  
the zebrafish embryo. *Development* **142**, 1050-1061, doi:10.1242/dev.119180  
(2015).
- 408 Elcheva, I., Brok-Volchanskaya, V. & Slukvin, I. Direct Induction of Hemogenic  
Endothelium and Blood by Overexpression of Transcription Factors in Human  
Pluripotent Stem Cells. *Journal of visualized experiments : JoVE*, e52910,  
doi:10.3791/52910 (2015).

- 409 Gerby, B. *et al.* SCL, LMO1 and Notch1 reprogram thymocytes into self-renewing cells. *PLoS Genet* **10**, e1004768, doi:10.1371/journal.pgen.1004768 (2014).
- 410 Kalender Atak, Z. *et al.* Comprehensive Analysis of Transcriptome Variation Uncovers Known and Novel Driver Events in T-Cell Acute Lymphoblastic Leukemia. *PLoS Genet* **9**, e1003997, doi:10.1371/journal.pgen.1003997 (2013).