I am still confused, but on a higher level.

Enrico Fermi
SUMMARY

The overall aim of the research described in this thesis was to improve our understanding of the mechanisms governing stem cell fate decisions and lineage commitment within the hematopoietic system. We tried to achieve this goal by exploiting two types of genetic variation in the mouse: naturally occurring genetic variation (chapters 2-6) and induced genetic variation (chapter 7). Here I summarize and discuss the work described in this thesis, and I conclude with an outlook on how the field may develop in the future.

Natural genetic variation

In part I of this thesis, we exploited naturally occurring genetic variation between the regular inbred mouse strains C57BL/6 (B6) and DBA/2 (D2) and across BXD (offspring B6 X D2) recombinant inbred mouse strains. We aimed to identify the genetic basis of variation in hematopoietic stem/progenitor cell (HSPC) traits, gene expression and microRNA expression, and to integrate the results. Changes in hematopoietic cell fate occur over time and are mediated by changes in regulatory networks. Yet, most studies of regulatory networks have provided “snapshots” usually at one potential regulatory level. In this thesis we analyzed network dynamics during the course of changes in hematopoietic cell fate, from Lin−Sca-1+c-Kit+ (LSK) multilineage cells to Lin−Sca-1−c-Kit+ progenitor cells to either TER-119+ erythroid cells or Gr-1+ myeloid cells. Also, we measured dynamic changes at multiple molecular levels. See Figure 1 for an illustrated summary of part I of this thesis.

In Chapter 2 we introduced the concepts of classical quantitative trait locus (QTL) mapping and expression QTL (eQTL) mapping. We reviewed past studies in which transcriptional profiling and/or genetic linkage analysis were performed on hematopoietic cells, and we discussed several potential future applications of eQTL mapping.

In Chapter 3 we performed an eQTL mapping study on four hematopoietic cell types isolated from the BXD mouse panel. This analysis allowed us to analyze eQTL robustness/sensitivity across different cellular differentiation states. We distinguished between “static” eQTLs that were consistently active in all four cell types and “dynamic” eQTLs that were sensitive to cell stage. Although we identified a large number (365) of static eQTLs, we identified a much larger number (1,283) of dynamic eQTLs. In total, we identified 140, 45, 531 and 295 eQTLs that were preferentially active in LSK, progenitor, erythroid, and myeloid cells, respectively. A detailed investigation of those dynamic eQTLs revealed that in the majority of cases the genes with a cell-type-specific eQTL were also most highly expressed in that particular cell type. We did not identify target genes that were regulated by distinct eQTLs in different cell types, suggesting that large-scale changes within functional regulatory networks are not common. Our
results demonstrate that the influence of genetic variants on gene expression is highly sensitive to the developmental stage of the cell population under study. Even when the purified cells were only separated by a few cell divisions, eQTLs demonstrated a remarkable plasticity. As the eQTL field is now moving to human (clinical) studies it is important to understand how eQTLs behave, and how they can be of use to study the etiology of disease. Therefore, it is imperative that future eQTL mapping studies are realized on multiple well-defined and highly purified cell types, ideally even on the scale of individual cells.

In **Chapter 4** we introduced a method that infers combinatorial association logic networks in multimodal genome-wide screens, and applied it to the myeloid gene expression dataset described in chapter 3. Unraveling (transcriptional) regulatory networks by inferring complex associations necessitates algorithms that take into account possible interactions. Therefore, instead of detecting direct associations between genetic loci and transcript levels (chapter 3), we detected associations between transcript levels and the outputs of small Boolean logic networks that combine multiple genetic loci. We identified 9 gene clusters that were significantly associated through a logical combination of genomic loci rather than a single eQTL. Notably, without incorporating interactions, these associations would have gone unnoticed.

In **Chapter 5** we exploited the naturally occurring genetic variation in the BXD mouse panel to construct gene networks that operate in successive stages of cellular development. We made use of the gene expression and eQTL datasets described in chapter 3. We first reconstructed an experimentally validated gene network, consisting of Gata1, Gata2, Pu.1, Tal1 and Zfpm1. We identified correlations between pairs of genes within this network, and superimposed these correlation measures on the predicted network model. Remarkably, we confirmed many of the proposed positive and negative interactions, either as positive or as negative correlations. Next, we constructed a novel gene network for those 52 transcripts that were most highly expressed in LSKs and down-regulated during both erythroid and myeloid differentiation. We used GeneNetwork to create an association network based on their pairwise correlation coefficients. Also, we extracted shared eQTLs and superimposed this information on the association network. We found that several transcripts were regulated by a region on chromosome 18, containing Zfp521, one of the 52 network members. Zfp521 itself was cis-regulated and therefore predicted to be upstream of the distantly co-regulated transcripts. Strikingly, the phenotypic trait hematopoietic stem cell (HSC) pool size had previously been mapped to this exact same locus, thereby pointing to Zfp521 as a candidate regulator of this trait. Functional studies on Zfp521 in HSPCs are ongoing.

In **Chapter 6** we determined whether variation in microRNA expression could be responsible for the observed variation in HSPC traits and gene expression across the BXD mouse panel. Therefore, we performed a genome-wide microRNA
expression study on four hematopoietic cell types isolated from the B6 and D2 parental mouse strains. We identified 131 microRNAs that were differentially expressed between cell types and 15 that were differentially expressed between mouse strains. Of special interest was an evolutionary conserved microRNA cluster located on chromosome 17 consisting of miR-99b, let-7e and miR-125a. All cluster members were most highly expressed in LSKs and down-regulated during both erythroid and myeloid differentiation. In addition, these microRNAs were higher expressed in D2 cells compared to B6 cells. To assess whether the differential expression of this microRNA cluster could be functional we overexpressed miR-99b, let-7e and miR-125a in primitive cells and quantified CFU-GM self-renewal, CAFC frequencies and LT-HSCs. In a way, we introduced genetic variation here, as we introduced extra copies of the microRNA cluster members. Bone marrow cells overexpressing the microRNA cluster showed increased replating capacity in CFU-GM assays and dramatically increased day-35 CAFC frequency. Furthermore, we found that mice reconstituted with these cells developed myeloproliferative neoplasms that occasionally progressed to leukemia. Finally, we identified 11 candidate functional downstream targets of the microRNA cluster.

**Induced genetic variation**

Instead of exploiting naturally occurring genetic variation, in part II of this thesis we introduced genetic variation into HSCs, with the aim to quantify the exact number of HSCs that actively contribute to hematopoiesis and to simultaneously analyze the behavior of multiple individual HSCs in a competitive polyclonal setting.

In Chapter 7 we introduced genetic variation in HSPCs by implementing a novel cellular barcoding technique. This technique makes use of retroviral plasmids that are labeled with random sequence tags or “barcodes”. Upon retroviral integration, each vector introduces a unique, identifiable and heritable mark into the host cell genome, allowing the clonal progeny of each cell to be tracked over time. In this chapter, we first validated our approach by performing clonal analyses on an immortalized cell line and a primary bone marrow culture. As expected, we found that the cell line retained its polyclonal nature over time, and that the primary cell culture exhibited decreased clonal complexity over time. Next, we used the barcoding technique to track hematopoietic repopulation dynamics in vivo after transplantation of barcoded HSPCs. Finally, we showed how clonal analysis can be complemented with transgene expression and integration site analysis. See Figure 1 for an illustrated summary of part II of this thesis.
Natural genetic variation

Genetic variation

C57BL/6

DBA/2

F1

F2 – F20

BXDs

Cell types

LSK

LS-K+

Ery

Mye

Variation in microRNA levels

Variation in mRNA levels

Variation in HSPC traits

Pool size

Proliferation rate

Reconstitution kinetics

Deterioration rate during aging

Mobilization response

Induced genetic variation

Clonal analysis

Clonal heterogeneity

Clonal dominance

Transgene expression analysis

Integration site analysis

Inverse PCR

LM-PCR

LAM-PCR

Figure 1. Illustrated summary of this thesis.
FUTURE PERSPECTIVES

Natural genetic variation
In this thesis we analyzed variation in phenotypic, cell biological and molecular traits in the classical inbred mouse strains C57BL/6 and DBA/2 and in recombinant inbred BXD mouse strains. Disadvantages of utilizing the BXD mouse panel are its limited genetic diversity and low number of recombinations (low mapping resolution). To overcome these limitations, future studies could utilize a “hybrid” strategy including classical inbred strains as well as recombinant inbred strains. The inbred strains would provide mapping resolution (fine-mapping) while the recombinant inbred strains would provide power. A mouse panel especially assembled for this purpose has been coined “Hybrid Mouse Diversity Panel (HMDP)” and consists of 100 commercially available inbred strains (i.e. 1 set of classical inbred strains, and 3 sets of recombinant inbred strains). These strains have all been fully genotyped and are renewable so that data can be collected ad infinitum. An alternative to the use of the HMDP would be to use outbred, heterogeneous stocks of mice. An advantage of using outbreds is that there is no limit to the number of genetically distinct animals, while the HMDP is limited to the number of available inbred strains. Disadvantages of using outbreds are the cost of high-density genotyping and the fact that they are not renewable.

Future studies could also utilize the Collaborative Cross (CC), a randomized cross of 8 inbred mouse strains: A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO, CAST/Ei, PWK/Ph, and WSB/Ei. The CC is characterized by high genetic diversity, large size (for high statistical power), and large number of recombinations (high mapping resolution). The target population size is 1,000 CC lines. This mouse panel will be more representative of the quantity of genetic variation present in human populations. By examining the phenotypic consequences of naturally occurring genetic variation in the mouse, insight into human complex diseases will be provided.

Many eQTL studies have reported the existence of eQTL hotspots (eQTLs regulating large numbers of transcripts). Interestingly, in chapter 3 we observed that these hotspots became more prevalent as differentiation from primitive to mature cell types ensued. Although the exact molecular nature of the hotspots remains unresolved at present, their increased potency as differentiation from stem cells to mature cells proceeds is compatible with the concept of multilineage priming of stem cells. Our data could suggest that relatively small transcriptional modules are active in primitive cells and that few, but potent, lineage specific regulators become activated during differentiation. However, this finding needs to be interpreted with caution, given the well-known volatility of eQTL hotspots.

Although eQTL hotspots have been documented in multiple studies it remains unclear by which molecular mechanism they emerge. In the simplest scenario they arise from polymorphisms in protein-coding genes that have multiple downstream
targets (e.g. transcription factor) or in microRNAs. Alternatively, eQTL hotspots may result from phenotypic buffering or widespread compensatory effects in response to local polymorphisms. It should be stressed that although a large collection of variable blood cell traits has been documented for the mouse strains analyzed in this thesis, they all display in essence normal blood cell production. The highly genotype-dependent gene expression patterns are therefore possibly compensating/buffering potential detrimental effects of naturally occurring polymorphisms, thereby ensuring normal blood cell production. Such a buffering system allows for highly stable gene networks and suggests substantial genetic redundancy.

Many challenges remain in the field of (e)QTL mapping. The past decade has been one of descriptive systems genetics studies. Many (e)QTLs have been identified, yet relatively few have been experimentally validated. Additionally, the subsequent identification of the polymorphisms that underlie those (e)QTLs is often hampered by large numbers of genes in the (e)QTL intervals and a high background of what is thought to be irrelevant sequence variation.

Induced genetic variation

In this thesis we validated a novel cellular barcoding technique for clonal analysis of complex cell populations in vitro and in vivo. We coupled the barcoding method to a Sanger sequencing-based detection system. A limitation of this approach is that the resolution by which minor clones in a complex population can be identified is restricted by the number of subclones sequenced. To overcome this limitation, ongoing studies in the lab focus on coupling the barcoding method to a high-throughput sequencing-based detection system. To make the high-throughput sequencing runs more cost-effective the lab also focuses on designing and validating multiplexing protocols that would allow the simultaneous analysis of tens to potentially hundreds of samples in a single sequencing run. In these protocols all barcoded samples are given unique identifiers by labeling them with different primers (“barcoding the barcode”). Preliminary results show that up to 75 samples could be simultaneously analyzed in a single Solexa sequencing run and that 4,000 to 2,000,000 sequence reads could be obtained per sample. This already allowed a quantitative, high-resolution assessment of clonal fluctuations in cultures of primary bone marrow cells. Future studies will focus on the detailed tracking of the behavior of multiple individual HSCs in a competitive polyclonal transplantation setting. Combining the barcoding method to a high-throughput sequencing-based detection system will offer a hitherto unprecedented sensitivity in analyzing cell population dynamics, making it especially suitable for detailed monitoring of gene-modified hematopoiesis in clinical gene therapy protocols.
Natural and induced genetic variation united

Global expression analyses have revealed that HSCs are “primed” for commitment by co-expression of multiple lineage-specific genes. As HSCs differentiate, “appropriate” lineage-specific genes are up-regulated, whereas “inappropriate” genes, specific for other lineages, are down-regulated. It is imperative that the gene expression changes that direct HSC fate decisions and lineage commitment are tightly controlled, as loss of this control results in hematopoietic disorders such as leukemias.

An interesting future study would be to perform a genetic analysis to identify leukemia-modulating genes. In such a study, both natural and induced genetic variation would be exploited to search for genes and gene networks that modify the progression of leukemia in mice. HSCs would be isolated from a large collection of genetically distinct mouse strains that have been densely genotyped. These HSCs would be transduced with a retroviral vector containing a (mild) oncogene and a barcode, and subsequently transplanted into recipients. The presence of the oncogene would make that these mice develop hematologic malignancies, and the presence of the barcode would make that the disease progression and type can be closely monitored over time. In addition, the barcode would allow transgene expression and integration site analysis. The transplanted mice would be carefully monitored and leukemic cell populations would be characterized and purified. Genome-wide studies would subsequently be implemented to detect changes in mRNA, microRNA and/or epigenetic levels. Finally, the relationships between genotype, molecular phenotypes and cell biological phenotypes would be analyzed.

Such a project would result in 1) the identification of genomic loci/genes that modulate (enhance or suppress) the development of leukemia, which may provide targets for therapy, 2) the generation of leukemic gene networks and how they associate with disease, which will contribute to the generation and validation of predictor profiles, and 3) the identification of haplotype-specific retroviral integration sites.