

**EXPLOITING NATURAL AND INDUCED GENETIC VARIATION
TO STUDY HEMATOPOIESIS**

ALICE GERRITS

The research described in this thesis was conducted at the Department of Cell Biology, Section Stem Cell Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.

The work described in this thesis was financially supported by:

- Graduate School for Drug Exploration (GUIDE)
- EUrythron, European Union, FP6
- EuroSyStem, European Union, FP7
- Dutch Cancer Society

The printing of this thesis was financially supported by:

- MPN Stichting (Nederlandse patiëntenorganisatie voor mensen met een myeloproliferatieve aandoening)
- Department of Cell Biology, Section Stem Cell Biology, University Medical Center Groningen, University of Groningen
- University Medical Center Groningen
- University of Groningen
- Graduate School for Drug Exploration (GUIDE)
- BD Biosciences
- Beckman Coulter B.V.

ISBN (printed version): 978-94-6182-002-0

ISBN (digital version): 978-94-6182-005-1

Lay-out and production: Off Page, Amsterdam, www.offpage.nl

Cover design: Off Page & A. Gerrits

© Copyright 2011 by A. Gerrits. All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, without prior permission of the author.



RIJKSUNIVERSITEIT GRONINGEN

**EXPLOITING NATURAL AND INDUCED GENETIC VARIATION
TO STUDY HEMATOPOIESIS**

Proefschrift

ter verkrijging van het doctoraat in de
Medische Wetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
woensdag 7 september 2011
om 16.15 uur

door

Alice Gerrits

geboren op 29 juli 1981
te Hardenberg

Promotor:	Prof. dr. G. de Haan
Copromotor:	Dr. L.V. Bystrykh
Beoordelingscommissie:	Prof. dr. C. Wijmenga Prof. dr. J.N.J. Philipsen Prof. dr. I. Röder

VOOR MIJN OUDERS

Paranimfen:

Ellen Weersing
Lisette Bok

CONTENTS

Chapter 1	Introduction & outline of the thesis	9
Chapter 2	Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny <i>Immunogenetics. 2008 Aug;60(8):411-22. Review</i>	21
Chapter 3	Expression quantitative trait loci are highly sensitive to cellular differentiation state <i>PLoS Genetics. 2009 Oct;5(10):e1000692</i> <i>Highlight in Nature Rev Genet. 2009 10 (12):819</i>	43
Chapter 4	Inferring combinatorial association logic networks in multimodal genome-wide screens <i>Bioinformatics. 2010 Jun 15;26(12):1149-57</i>	61
Chapter 5	Genetic screen identifies <i>Zfp521</i> as a candidate regulator of hematopoietic stem cell pool size <i>In preparation</i>	85
Chapter 6	Genetic screen identifies microRNA cluster 99b/ <i>let-7e/125a</i> as a regulator of primitive hematopoietic cells <i>Invited for resubmission to Blood</i>	99
Chapter 7	Cellular barcoding tool for clonal analysis in the hematopoietic system <i>Blood. 2010 Apr 1;115(13):2610-8</i> <i>Appeared on the cover of Blood</i>	123
Chapter 8	Summarizing discussion	145
Appendices	Contributing authors	157
	Nederlandse samenvatting (voor niet-ingewijden)	161
	Dankwoord	167
	Biografie	175
	Biography	176
	List of publications	177
	Attended meetings	178
	Funding and awards	179

CHAPTER 1

INTRODUCTION & OUTLINE OF THE THESIS

Begin at the beginning and go on till you come to the end: then stop.

The King, Lewis Carroll's Alice in Wonderland

INTRODUCTION

The Hematopoietic System

The blood is one of the most intensely studied and best understood human tissues. It performs many functions in the body and consists of red blood cells, white blood cells and platelets suspended in plasma. Red blood cells (also referred to as erythrocytes) are the most common type of blood cell and carry oxygen to the body tissues. White blood cells are much less numerous than red blood cells. Their function is to protect the body from infection. White blood cells consist of myeloid and lymphoid cells. The myeloid cell compartment is made up of granulocytes (neutrophils, basophils, eosinophils) and monocytes/macrophages that collectively combat infections from bacteria, fungi, and other parasites such as worms. Some of these cells also have the ability to remove dead or damaged tissues. The lymphoid cell compartment is made up of B- and T-lymphocytes. B-lymphocytes produce antibodies that bind to antigens, while T-lymphocytes can directly attack and destroy cells recognized as foreign to the body, including virus-infected cells and cancer cells. Platelets (also referred to as thrombocytes) are derived from fragmentation of megakaryocytes and play an essential role in the blood clotting process.

Many blood cells are short-lived and therefore must be constantly replenished. A healthy human adult requires billions of new blood cells on a daily basis. The process by which these cells are produced is called hematopoiesis and is maintained by a small population of hematopoietic stem cells (HSCs).

Hallmark Properties of HSCs

The hematopoietic system is hierarchically organized with HSCs at the top of the pyramid and mature blood cells at the bottom of the pyramid. HSCs have two core properties, the combination of which makes them unique. First, they have long-term self-renewal potential, meaning that they can make identical copies of themselves for long periods of time. Second, they have multi-lineage differentiation potential, meaning that they can differentiate into the many specialized cell types that make up the blood system. HSCs generate intermediate cell types or progenitors that proliferate and become progressively more restricted in developmental potential, until ultimately mature blood cells are formed (Figure 1). Careful control of the balance between HSC self-renewal and differentiation is essential to maintain blood cell homeostasis. In a homeostatic system, the HSC self-renewal probability P is 0.5, indicating that on average half of the progeny of the HSCs preserves stem cell characteristics, whereas the other half initiates differentiation. If self-renewal of HSCs is impaired ($P < 0.5$) the stem cell pool will exhaust, ultimately resulting in bone marrow failure. In contrast, if self-renewal of HSCs is increased ($P > 0.5$) the stem cell pool will expand, ultimately resulting in blood cancer.¹

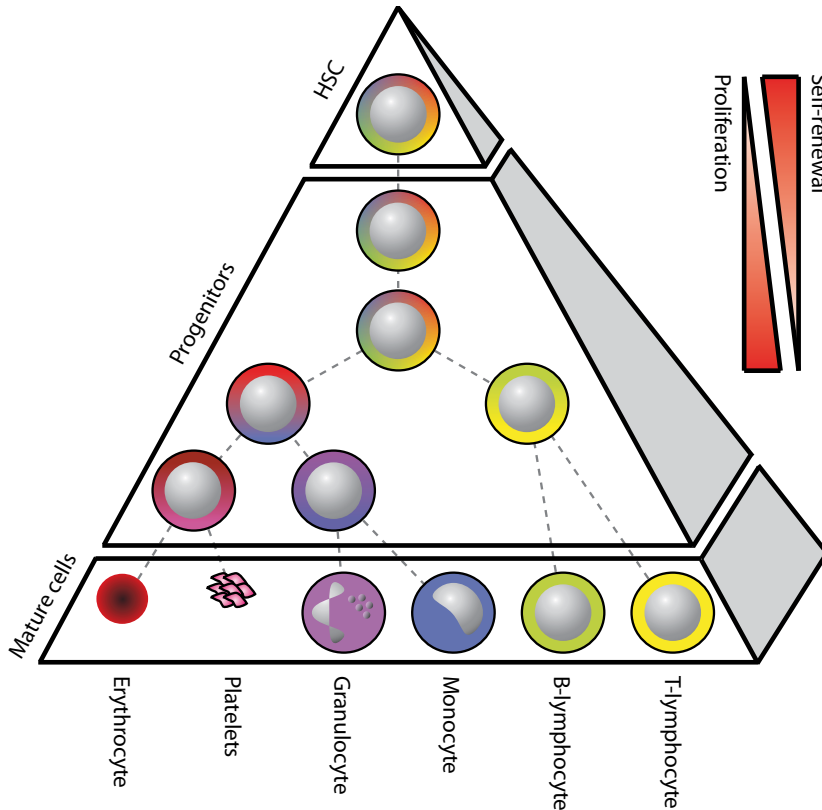


Figure 1. Hierarchical organization of the blood system. Hematopoietic stem cells (HSCs) give rise to more committed progenitors that in turn produce the different blood cell types.

HSCs are forced to constantly make fate decisions; they may remain quiescent, undergo apoptosis or initiate cell division. Once the decision to divide has been made, they must choose whether to self-renew or to differentiate into a specific blood cell type. HSC fate decisions are controlled by a complex interplay between cell-autonomous (intrinsic) and cell-nonautonomous (extrinsic) signals. Yet, the exact mechanisms governing these decisions are still poorly understood. The continued quest for these mechanisms will not only be of major relevance for a fundamental understanding of normal hematopoiesis and developmental biology, but also for a better understanding of the pathogenesis and evolution of blood cancers. In addition, it will open new avenues for the design of *ex vivo* expansion protocols for HSCs, thereby increasing their clinical potential in cell replacement and gene therapy protocols. The delineation of the mechanisms underlying HSC fate decisions requires the phenotypic, functional and molecular characterization of HSCs.

Phenotypic Characterization of HSCs

Characterizing HSCs began by studies in rodents, which laid the foundation for studies in humans. The major challenge in HSC studies is that the stem cells themselves are rare, representing less than 0.01% of the cells within the bone marrow, and that the stem cells cannot be identified based on their morphology (size and shape), but solely based on their immunophenotype. The general consensus is that mouse HSCs lack the cell surface markers that are expressed on differentiated hematopoietic cells (so-called lineage or lin markers), but do express "stem cell antigen-1" (Sca-1) and c-Kit (collectively referred to as LSK).^{2;3} The cells in this LSK population can be further subfractionated into long-term (LT) and short-term (ST) HSCs (both containing self-renewal capacity) and progenitors (containing low/no self-renewal capacity). LT-HSCs can be more precisely identified within the LSK population by selecting for those cells that are CD34⁻,⁴ Flk-2/Flt3⁻,^{5;6} CD48⁻ and CD150⁺,⁷ EPCR/CD201⁺,⁸ and different combinations thereof. Another approach to identify HSCs relies on their ability to efflux vital dyes such as Hoechst 33342⁹ and Rhodamine 123.¹⁰⁻¹³ The prospective isolation of mouse HSCs by fluorescence-activated cell sorting (FACS) has facilitated the functional and molecular characterization of HSCs.

Functional Characterization of HSCs

HSC studies depend on several *in vitro* and *in vivo* assays to measure the quantity and/or quality of hematopoietic stem and progenitor cells. Several *in vitro* assays have been established that can measure the frequency of progenitors (colony-forming unit in culture; CFU-C), stem cells (long-term culture-initiating cell; LTC-IC) or both (cobblestone-area forming cell assay; CAFC).^{14;15} At present, the ultimate model used to demonstrate HSC activity is *in vivo* transplantation; the only assay that can measure both the long-term self-renewal and multi-lineage differentiation capacity of HSCs.^{16;17} In this assay, HSCs are transplanted into myeloablated (usually irradiated) recipients after which their ability to produce multi-lineage progeny for prolonged periods of time is evaluated. It should be realized, however, that in all the above-mentioned assays HSCs are measured retrospectively. By the time HSC activity is established, the original HSC will unavoidably be lost.

The existence of true HSCs with long-term self-renewal activity and multi-lineage differentiation capacity has been irrefutably demonstrated by retroviral marking studies,^{18;19} limiting dilution²⁰⁻²² and single purified cell assays.²³⁻²⁶ However, thus far it has remained difficult to quantify the exact number of HSCs that actively contribute to hematopoiesis or to simultaneously analyze the behavior of multiple individual HSCs in a competitive polyclonal setting. That type of analysis could help resolve some of the major outstanding questions in HSC biology, such as whether lineage-biased and dormant or hibernating HSCs really do exist,²⁷⁻³⁰ and how clonal relationships in the hematopoietic system are established.

Molecular Characterization of HSCs

Systems Biology

State-of-the-art genome-wide tools have become available to elucidate the intrinsic properties of HSCs. To identify genes that play a key role in regulating HSC fate decisions, several groups have embarked on genome-wide mRNA expression studies. In an attempt to identify a HSC molecular signature, the expression profiles of HSCs, embryonic stem cells and neural stem cells have been compared.^{31;32} Furthermore, mRNA profiles have been analyzed in HSCs from young versus old animals,^{33;34} in quiescent versus proliferating HSCs,³⁵ and in HSCs versus their differentiated progeny.³⁶⁻³⁹ However, to explain complex biological phenomena it is not enough to study mRNA expression alone; it is of vital importance to measure – in the same sample – complementary biological variables. Only then HSCs can be studied from a systems perspective.

More recently, novel genome-wide tools have been developed to study alternative splicing, microRNA expression, protein expression and post-translational modifications, transcription factor and cofactor binding sites, epigenetic marks such as histone modifications and DNA methylation, and chromatin accessibility. All these types of analyses thus far remain challenging in HSCs, because of the limiting cell numbers available per mouse. Yet, technology is rapidly evolving and may soon provide a hitherto unprecedented amount of data on HSCs. The challenge that therefore lies ahead is to integrate all these different types of data using various bioinformatics and systems biology approaches. The ultimate goal will be to construct regulatory networks that allow us to model, predict and potentially control HSC fate.

Systems Genetics

Quantitative differences can be observed for most complex behavioral and physiological traits among individuals within any population. These differences can be due to both genetic and environmental factors. The mouse, with its long history of genetic research, has proven to be an excellent model to study the genetic basis for these quantitative differences. Specifically, the regular inbred mouse strains C57BL/6 (B6) and DBA/2 (D2) have been extensively studied. With respect to HSCs these strains differ for example in their pool size, proliferation rates, reconstitution kinetics, deterioration rate during aging and mobilization response.⁴⁰⁻⁴⁶ The underlying causes of variation in these quantitative traits must be encoded in the genomes of these mice.

The use of the BXD recombinant inbred mouse panel has proven to be a useful strategy to pinpoint the genomic regions that could be responsible for the observed variation in HSC traits. This panel has been derived by crossing B6 and D2 mouse strains and then inbreeding progeny for many generations.⁴⁷⁻⁴⁹ As a result, the genomes of both parental strains have recombined and have become fixed (i.e. are homozygous) in a unique pattern in each of the BXD lines. The panel

of BXDs has been extensively genotyped and phenotyped. In order to identify the genomic regions that influence the above-mentioned HSC traits an approach called quantitative trait locus (QTL) mapping could be employed. This approach makes use of statistical methods that search for associations between variation in the phenotypic trait and possession of a particular allele. In this manner, genomic loci have been identified that influenced the pool size, proliferation rate and mobilization response of HSCs.⁵⁰⁻⁵⁵

More recently, an approach called expression QTL (eQTL) mapping has been employed to delineate the genetic basis of gene expression. In this approach, variation in transcript abundance is considered to be a quantitative trait that can be mapped to genomic loci.^{56;57} This approach has been implemented by our laboratory on a genome-wide scale to study gene regulation in HSCs.⁵⁸ Soon after, a combinatorial analysis of physiological QTLs and eQTLs led to the identification of *Lxn* as a gene involved in regulating HSC pool size,⁵⁹ and EGFR signaling as a pathway involved in regulating HSC mobilization.⁶⁰ Discoveries such as these have invigorated the field of HSC biology and have fuelled excitement to continue systems genetics studies on HSCs.

OUTLINE OF THE THESIS

The overall aim of the research described in this thesis is to improve our understanding of the mechanisms governing stem cell fate decisions and lineage commitment within the hematopoietic system. We try to achieve this goal by exploiting two types of genetic variation in the mouse:

- I) Naturally occurring genetic variation (chapters 2-6)
- II) Induced genetic variation (chapter 7)

Chapter 2 introduces the concepts of classical QTL mapping and expression QTL mapping. It reviews past studies in which transcriptional profiling and/or genetic linkage analysis were performed on hematopoietic cells, and serves as an introduction to chapters 3-6.

Chapter 3 describes an eQTL mapping study on four hematopoietic cell types isolated from a panel of BXD recombinant inbred mouse strains. Thus far, most eQTL mapping studies that have been reported have analyzed single cell types or compared developmentally unrelated and distant cell types. This chapter reports the first study of eQTL dynamics across closely related cell types during cellular development. It covers the identification of consistently active (or “static”) eQTLs and cell-type-dependent (or “dynamic”) eQTLs.

Chapter 4 introduces a method that infers combinatorial association logic networks in multimodal genome-wide screens. Traditional eQTL mapping methods detect direct associations between genetic loci and transcript levels.

However, when the genetic loci themselves interact, direct associations between the individual loci and transcript levels may become undetectable. To alleviate this problem, the method described in this chapter detects associations between transcript levels and the outputs of small Boolean logic networks that combine multiple genetic loci.

Chapter 5 illustrates how the presence of genetic variation in a pedigree of normal, wild-type, mice can be exploited to (re)construct gene networks that operate in successive stages of cellular development. It starts with the reconstruction of an experimentally validated gene network, and proceeds with the construction of a novel gene network for transcripts that are specific to the HSC-enriched cell population. Finally, a combinatorial analysis of classical QTL and eQTL data leads to the identification of a zinc-finger transcription factor as a candidate regulator of the size of the HSC pool.

Chapter 6 describes a microRNA profiling study on four developmentally related hematopoietic cell types isolated from B6 and D2 mouse strains. It covers the identification of cell type-dependent and mouse strain-dependent microRNAs. Of special interest is an evolutionary conserved microRNA cluster that is most highly expressed in the HSC-enriched cell population and that is differentially expressed between mouse strains. To assess whether the differential expression of this cluster could be functional, we overexpressed it beyond its normal expression level, and found that this fixed hematopoietic cells in a primitive state. Finally, we identify the downstream mRNA targets through which this microRNA cluster may exert its effect.

Chapter 7 introduces a novel cellular barcoding technique that can be used as a powerful tool to assay clonality in the hematopoietic system. 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. We demonstrate the efficacy of the barcoding technique to track clonal dynamics in two distinct cell culture systems *in vitro* and in a hematopoietic transplantation setting *in vivo*, and emphasizes the importance of implementing barcoded vectors in all future clinical gene therapy protocols.

Chapter 8 summarizes and discusses the results obtained in the previous chapters.

REFERENCES

1. De Haan G, Gerrits A. Epigenetic control of hematopoietic stem cell aging the case of Ezh2. *Ann.N.Y.Acad.Sci.* 2007;1106233-239.
2. Spangrude GJ, Heimfeld S, Weissman IL. Purification and characterization of mouse hematopoietic stem cells. *Science* 1988;241(4861):58-62.
3. Okada S, Nakauchi H, Nagayoshi K et al. In vivo and in vitro stem cell function of c-kit and Sca-1-positive murine hematopoietic cells. *Blood* 1992;80(12):3044-3050.
4. Osawa M, Hanada K, Hamada H, Nakauchi H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* 1996;273(5272):242-245.
5. Christensen JL, Weissman IL. Flk-2 is a marker in hematopoietic stem cell differentiation: a simple method to isolate long-term stem cells. *Proc.Natl.Acad.Sci.U.S.A* 2001;98(25):14541-14546.
6. Adolfsson J, Borge OJ, Bryder D et al. Upregulation of Flt3 expression within the bone marrow Lin(-)Sca1(+)c-kit(+) stem cell compartment is accompanied by loss of self-renewal capacity. *Immunity*. 2001;15(4):659-669.
7. Kiel MJ, Yilmaz OH, Iwashita T et al. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* 2005;121(7):1109-1121.
8. Balazs AB, Fabian AJ, Esmon CT, Mulligan RC. Endothelial protein C receptor (CD201) explicitly identifies hematopoietic stem cells in murine bone marrow. *Blood* 2006;107(6):2317-2321.
9. Goodell MA, Brose K, Paradis G, Conner AS, Mulligan RC. Isolation and functional properties of murine hematopoietic stem cells that are replicating in vivo. *J.Exp.Med.* 1996;183(4):1797-1806.
10. Mulder AH, Visser JW. Separation and functional analysis of bone marrow cells separated by rhodamine-123 fluorescence. *Exp.Hematol.* 1987;15(1):99-104.
11. Li CL, Johnson GR. Rhodamine123 reveals heterogeneity within murine Lin-, Sca-1+ hemopoietic stem cells. *J.Exp.Med.* 1992;175(6):1443-1447.
12. Bertocello I, Hodgson GS, Bradley TR. Multiparameter analysis of transplantable hemopoietic stem cells: I. The separation and enrichment of stem cells homing to marrow and spleen on the basis of rhodamine-123 fluorescence. *Exp.Hematol.* 1985;13(10):999-1006.
13. Spangrude GJ, Johnson GR. Resting and activated subsets of mouse multipotent hematopoietic stem cells. *Proc.Natl.Acad.Sci.U.S.A* 1990;87(19):7433-7437.
14. Van Os RP, Dethmers-Ausema B, De Haan G. In vitro assays for cobblestone area-forming cells, LTC-IC, and CFU-C. *Meth-ods Mol.Biol.* 2008;430143-157.
15. Ploemacher RE, Van der Sluijs JP, Van Beurden CA, Baert MR, Chan PL. Use of limiting-dilution type long-term marrow cultures in frequency analysis of marrow-repopulating and spleen colony-forming hematopoietic stem cells in the mouse. *Blood* 1991;78(10):2527-2533.
16. Ford CE, Hamerton JL, Barnes DW, Loutit JF. Cytological identification of radiation-chimaeras. *Nature* 1956;177(4506):452-454.
17. McCulloch EA, Till JE. The radiation sensitivity of normal mouse bone marrow cells, determined by quantitative marrow transplantation into irradiated mice. *Radiat. Res.* 1960;13115-125.
18. Capel B, Hawley RG, Mintz B. Long- and short-lived murine hematopoietic stem cell clones individually identified with retroviral integration markers. *Blood* 1990;75(12):2267-2270.
19. Jordan CT, Lemischka IR. Clonal and systemic analysis of long-term hematopoiesis in the mouse. *Genes Dev.* 1990;4(2):220-232.
20. Smith LG, Weissman IL, Heimfeld S. Clonal analysis of hematopoietic stem-cell differentiation in vivo. *Proc.Natl.Acad.Sci.U.S.A* 1991;88(7):2788-2792.
21. Zhong RK, Astle CM, Harrison DE. Distinct developmental patterns of short-term and long-term functioning lymphoid and myeloid precursors defined by competitive limiting dilution analysis in vivo. *J.Immunol.* 1996;157(1):138-145.
22. Cho RH, Muller-Sieburg CE. High frequency of long-term culture-initiating cells retain in vivo repopulation and self-renewal capacity. *Exp.Hematol.* 2000;28(9):1080-1086.
23. Takano H, Ema H, Sudo K, Nakauchi H. Asymmetric division and lineage commitment at the level of hematopoietic stem cells: inference from differentiation in daughter cell and granddaughter cell pairs. *J.Exp.Med.* 2004;199(3):295-302.

24. Ema H, Sudo K, Seita J et al. Quantification of self-renewal capacity in single hematopoietic stem cells from normal and Lnk-deficient mice. *Dev.Cell* 2005;8(6):907-914.
25. Dykstra B, Kent D, Bowie M et al. Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell* 2007;1(2):218-229.
26. Osawa M, Hanada K, Hamada H, Nakauchi H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* 1996;273(5272):242-245.
27. Muller-Sieburg CE, Cho RH, Thoman M, Adkins B, Sieburg HB. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood* 2002;100(4):1302-1309.
28. Yamazaki S, Iwama A, Takayanagi S et al. Cytokine signals modulated via lipid rafts mimic niche signals and induce hibernation in hematopoietic stem cells. *EMBO J.* 2006;25(15):3515-3523.
29. Wilson A, Laurenti E, Oser G et al. Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* 2008;135(6):1118-1129.
30. Foudi A, Hochedlinger K, Van Buren D et al. Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. *Nat.Biotechnol.* 2009;27(1):84-90.
31. Ivanova NB, Dimos JT, Schaniel C et al. A stem cell molecular signature. *Science* 2002;298(5593):601-604.
32. Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA. "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* 2002;298(5593):597-600.
33. Rossi DJ, Bryder D, Zahn JM et al. Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proc.Natl.Acad.Sci.U.S.A* 2005;102(26):9194-9199.
34. Chambers SM, Shaw CA, Gatza C et al. Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation. *PLoS.Biol.* 2007;5(8):e201.
35. Venezia TA, Merchant AA, Ramos CA et al. Molecular signatures of proliferation and quiescence in hematopoietic stem cells. *PLoS.Biol.* 2004;2(10):e301.
36. Forsberg EC, Prohaska SS, Katzman S et al. Differential expression of novel potential regulators in hematopoietic stem cells. *PLoS.Genet.* 2005;1(3):e28.
37. Kiel MJ, Yilmaz OH, Iwashita T et al. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* 2005;121(7):1109-1121.
38. Chambers SM, Boles NC, Lin KY et al. Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell* 2007;1(5):578-591.
39. Terskikh AV, Miyamoto T, Chang C, Diatchenko L, Weissman IL. Gene expression analysis of purified hematopoietic stem cells and committed progenitors. *Blood* 2003;102(1):94-101.
40. De Haan G, Nijhof W, Van Zant G. Mouse strain-dependent changes in frequency and proliferation of hematopoietic stem cells during aging: correlation between lifespan and cycling activity. *Blood* 1997;89(5):1543-1550.
41. De Haan G, Van Zant G. Intrinsic and extrinsic control of hemopoietic stem cell numbers: mapping of a stem cell gene. *J.Exp.Med.* 1997;186(4):529-536.
42. Van Zant G, Eldridge PW, Behringer RR, Dewey MJ. Genetic control of hematopoietic kinetics revealed by analyses of allophenic mice and stem cell suicide. *Cell* 1983;35(3 Pt 2):639-645.
43. Muller-Sieburg CE, Riblet R. Genetic control of the frequency of hematopoietic stem cells in mice: mapping of a candidate locus to chromosome 1. *J.Exp.Med.* 1996;183(3):1141-1150.
44. Hasegawa M, Baldwin TM, Metcalf D, Foote SJ. Progenitor cell mobilization by granulocyte colony-stimulating factor controlled by loci on chromosomes 2 and 11. *Blood* 2000;95(5):1872-1874.
45. Roberts AW, Foote S, Alexander WS et al. Genetic influences determining progenitor cell mobilization and leukocytosis induced by granulocyte colony-stimulating factor. *Blood* 1997;89(8):2736-2744.
46. De Haan G, Szilvassy SJ, Meyerrose TE et al. Distinct functional properties of highly purified hematopoietic stem cells from mouse strains differing in stem cell numbers. *Blood* 2000;96(4):1374-1379.
47. Bailey DW. Recombinant-inbred strains. An aid to finding identity, linkage, and function of histocompatibility and other genes. *Transplantation* 1971;11(3):325-327.
48. Taylor, B. A. Recombinant inbred strains: use in gene mapping. In: Morse III HC (ed), *Origins of Inbred Mice*. Academic Press, NY. 423-438. 1978.

49. Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC.Genet.* 2004;57.
50. Muller-Sieburg CE, Riblet R. Genetic control of the frequency of hematopoietic stem cells in mice: mapping of a candidate locus to chromosome 1. *J.Exp.Med.* 1996;183(3):1141-1150.
51. De Haan G, Bystrykh LV, Weersing E et al. A genetic and genomic analysis identifies a cluster of genes associated with hematopoietic cell turnover. *Blood* 2002;100(6):2056-2062.
52. De Haan G, Van Zant G. Genetic analysis of hemopoietic cell cycling in mice suggests its involvement in organismal life span. *FASEB J.* 1999;13(6):707-713.
53. Geiger H, True JM, De Haan G, Van Zant G. Age- and stage-specific regulation patterns in the hematopoietic stem cell hierarchy. *Blood* 2001;98(10):2966-2972.
54. De Haan G, Van Zant G. Intrinsic and extrinsic control of hemopoietic stem cell numbers: mapping of a stem cell gene. *J.Exp.Med.* 1997;186(4):529-536.
55. Geiger H, Szilvassy SJ, Ragland P, Van Zant G. Genetic analysis of progenitor cell mobilization by granulocyte colony-stimulating factor: verification and mechanisms for loci on murine chromosomes 2 and 11. *Exp.Hematol.* 2004;32(1):60-67.
56. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Genet.* 2001;17(7):388-391.
57. Schadt EE, Monks SA, Drake TA et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 2003;422(6929):297-302.
58. Bystrykh L, Weersing E, Dontje B et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat.Genet.* 2005;37(3):225-232.
59. Liang Y, Jansen M, Aronow B, Geiger H, Van Zant G. The quantitative trait gene latexin influences the size of the hematopoietic stem cell population in mice. *Nat.Genet.* 2007;39(2):178-188.
60. Ryan MA, Nattamai KJ, Xing E et al. Pharmacological inhibition of EGFR signaling enhances G-CSF-induced hematopoietic stem cell mobilization. *Nat.Med.* 2010;16(10):1141-1146.

CHAPTER 2

COMBINING TRANSCRIPTIONAL PROFILING AND GENETIC LINKAGE ANALYSIS TO UNCOVER GENE NETWORKS OPERATING IN HEMATOPOIETIC STEM CELLS AND THEIR PROGENY

Alice Gerrits, Brad Dykstra, Marcel Otten,
Leonid V. Bystrykh and Gerald de Haan

Immunogenetics. 2008 Aug;60(8):411-22. Review

ABSTRACT

Stem cells are unique in that they possess both the capacity to self-renew and thereby maintain their original pool, as well as the capacity to differentiate into mature cells. In the past number of years, transcriptional profiling of enriched stem cell populations has been extensively performed in an attempt to identify a universal stem cell gene expression signature. While stem cell-specific transcripts were identified in each case, this approach has thus far been insufficient to identify a universal group of core “stemness” genes ultimately responsible for self-renewal and multipotency. Similarly, in the hematopoietic system, comparisons of transcriptional profiles between different hematopoietic cell stages have had limited success in revealing core genes ultimately responsible for the initiation of differentiation and lineage specification. Here, we propose that the combined use of transcriptional profiling and genetic linkage analysis, an approach called “genetical genomics”, can be a valuable tool to assist in the identification of genes and gene networks that specify “stemness” and cell fate decisions. We review past studies of hematopoietic cells that utilized transcriptional profiling and/or genetic linkage analysis, and discuss several potential future applications of genetical genomics.

INTRODUCTION

Stem cells are defined by their capacity to self-renew and their ability to differentiate into mature cell types. Stem cells can be divided into two main categories: embryonic stem cells (ESCs) and “tissue-specific” stem cells. ESCs, which are derived from the blastocyst stage of the early embryo, are termed pluripotent because they are able to differentiate into cells of all three germ layers; ectoderm, endoderm and mesoderm. In contrast, tissue-specific stem cells are termed multipotent because they are only able to differentiate into a limited number of more closely related cell types. Tissue-specific stem cells have now been identified in a number of fetal and adult tissues, where they play essential roles in organogenesis, tissue homeostasis and repair.

One of the major challenges in the stem cell field has been to identify a universal “stem cell signature”, that is, those genes that ultimately enable the hallmark stem cell features of self-renewal and pluripotency or multipotency. In addition, it is important to identify those genes that initiate differentiation and determine lineage specification. In this review, we discuss how these challenges might be addressed using the combinatorial approach of genetical genomics. In particular, we focus on hematopoietic stem cells (HSCs), one of the best studied stem cell systems, as a model to investigate cell fate decisions.

HSCs are rare cells in the bone marrow that both self-renew and generate differentiated blood cells. During the process of hematopoietic differentiation, the cells progressively amplify their numbers, lose their multipotency and become increasingly committed. Ultimately, HSCs are able to give rise to large numbers of cells of both myeloid (e.g. monocytes/ macrophages, granulocytes, erythrocytes and megakaryocytes/ platelets) and lymphoid lineages (e.g. T cells and B cells).¹ A simplified overview of hematopoiesis is depicted in Figure 1. Although HSCs, intermediate progenitors, and most mature blood cells are genetically identical and are only a few cell divisions apart from each other, they differ tremendously in both phenotype and function.

In search of a stem cell signature

Since HSCs share certain hallmark properties with other stem cell types, it was speculated that a universal stem cell signature, consisting of a common set of genes whose concerted expression grant stem cells their unique properties, might exist. In an attempt to identify such a signature, the expression profiles of mouse ESCs, HSCs and neural stem cells (NSCs) were compared by two groups independently.^{2,3} Both studies revealed similarities and differences between these cell populations, and generated a list of over 200 stem cell-specific genes that mouse ESCs, HSCs and NSCs commonly expressed. Strikingly however, these two lists shared only six genes. Soon thereafter, a third independent expression profiling study comparing ESCs, NSCs and retinal progenitor/ stem cells was reported.

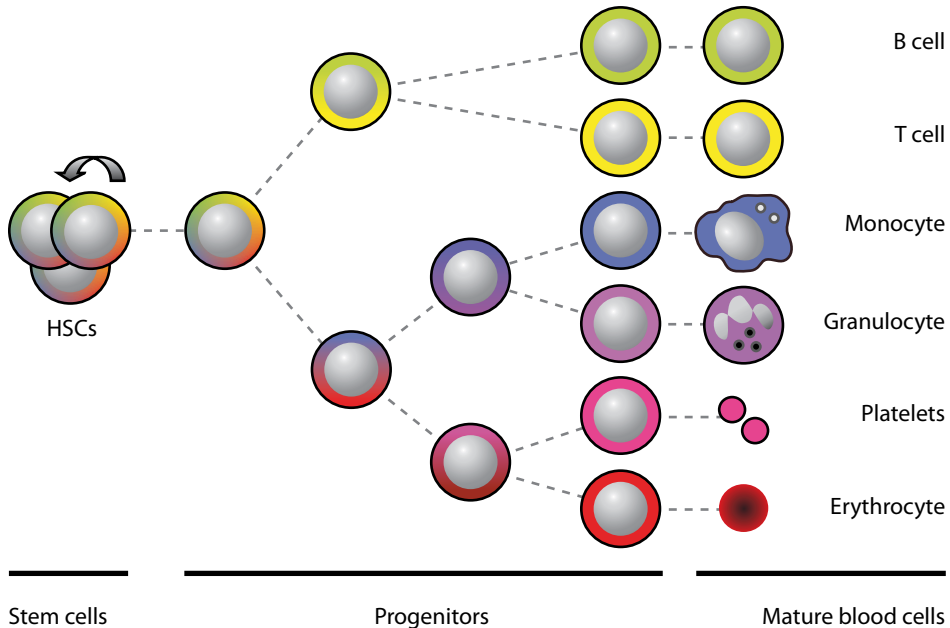


Figure 1. Simplified overview of hematopoiesis. Hematopoietic stem cells (HSCs) have self-renewal activity (represented as the *arrow*) and can therefore maintain their numbers. During hematopoietic differentiation, HSCs lose their self-renewal capacity and become increasingly lineage-committed (represented as the gradual loss of colors).

Comparing these three independently generated lists of “stemness” genes, only one gene (*Itga6*) was commonly identified.⁴ What could explain the lack of overlap between these lists? First, differences in methodology may be in part responsible (discussed in ^{5;6}). Second, ESCs are pluripotent, while the other stem cell types in the described comparisons are more restricted in their developmental potential. These different stem cell populations were compared in a direct manner, although it is certainly possible that pluripotency and multipotency are maintained by different gene circuits. A third, more fundamental possibility is that a universal stem cell signature simply does not exist. ESCs, HSCs, NSCs and retinal stem cells may each have their very own transcriptional networks responsible for their unique stem cell properties. If this is the case, cross-tissue comparative stem cell transcriptome analyses are not particularly useful approaches to identify these stem cell type-specific signatures.

That a limited number of factors could indeed specify “stemness” was convincingly shown by Takahashi and Yamanaka, who demonstrated that the forced expression of Oct4, Sox2, Klf4 and c-Myc and could reprogram mouse fibroblasts into pluripotent stem cells that were functionally equivalent to ESCs.⁷ Subsequently, it was shown that some family proteins of the four factors could also reprogram fibroblasts, and that c-Myc was dispensable for this process.^{8;9}

The question whether terminally differentiated cells could also be reprogrammed was resolved when fully differentiated mature mouse B lymphocytes were reprogrammed to pluripotency.¹⁰ Induction of pluripotency was also recently demonstrated in human fibroblasts using various combinations of factors.¹¹⁻¹³ The demonstration that more than one combination of factors could confer the same stem cell characteristics suggests that stem cell signatures may be “degenerate” (as in the “*degenerate* DNA code”).

Thus, it is clear that “stemness” can be achieved by only a limited number of key stem cell regulators, presumably targeting larger collections of downstream genes in a hierarchical manner. Extracting such key regulators (or causes) from their downstream target genes (consequences) is not feasible using micro-array profiling approaches alone.

In search of hematopoietic fate determinants

Transcriptional profiling has also been utilized extensively in an attempt to identify genes whose expression distinguishes HSCs from their downstream progeny. Global expression analyses have revealed that stem cells exist in a “promiscuous” state where multiple lineage-specific genes are co-expressed, albeit at very low levels. Upon differentiation, “appropriate” lineage-specific genes are up-regulated, whereas “inappropriate” genes, specific for other lineages, are down-regulated.¹⁴ Recently, Chambers *et al* generated an expression database of various hematopoietic cell types, including HSCs, erythroid cells, granulocytes, monocytes, natural killer cells, activated and naive T cells, and B cells.¹⁵ This comparative transcriptome analysis provided large lists of genes that are specifically expressed in one cell stage or cell type compared to another. However, it is improbable that the transition from one cell stage to another relies on the independent regulation of so many genes. More likely, activation of a limited number of key regulatory genes initiates a cascade of events, resulting in the altered expression of tens to hundreds of genes.

Transcriptional profiling has proven to be a useful approach to identify cell stage and cell type-specific transcripts. When combined with other genetic approaches, it may also have the potential to identify key regulatory genes.

HSCs and linkage genetics

It has become clear that many hematopoietic characteristics or traits are genetically controlled, since they differ between various strains of genetically distinct laboratory mice. For example, a substantial strain-to-strain variation in the number of primitive hematopoietic cells and their turnover rates has been observed. Interestingly, an inverse correlation was detected between progenitor cell turnover rate and mouse lifespan.¹⁶

Two regular inbred strains of mice, C57BL/6 (B6) and DBA/2 (D2), have distinct differences in both their HSC traits and lifespan. Compared to B6 mice,

D2 mice have a shorter lifespan, a substantially higher HSC frequency, and their progenitors cycle at a much faster rate.¹⁶⁻¹⁹ In B6 mice the HSC frequency increases at a constant rate during the aging process,²⁰⁻²³ while in D2 mice it increases up to one year of age and then drops again.^{24,25} The observed natural variation between these regular inbred mouse strains offers a powerful tool to study the genetic basis of variation in these traits. The use of B6 x D2 (BXD) recombinant inbred mouse strains has been a particularly useful strategy to identify genomic regions affecting traits of interest. These inbred lines were developed by crossing the two inbred parental strains followed by repeated sibling-sibling mating for a minimum of 20 generations. The resulting BXD mouse strains each carry a genome that consists of a unique mosaic of homozygous B6 and D2 segments. At present, the BXD panel is composed of 80 different strains that all have been fully genotyped.²⁶ Variation in any quantifiable trait can be associated with the segregation of parental alleles, and linkage genetics can map this variation to quantitative trait loci (QTLs), thereby identifying the genomic region(s) affecting that trait. An overview of the QTL mapping approach is depicted in Figure 2.

Classical QTL analysis has permitted the identification of loci that are associated with variation in HSC traits. When HSC frequency was measured in the BXD reference panel using long-term culture initiating cell assays (LTC-ICs), two suggestive QTLs on chromosome 1 and one on chromosome 11 were identified. One of the loci on chromosome 1 was confirmed to affect HSC frequency in a congenic mouse strain.¹⁷ When HSC frequency was assessed using cobblestone area forming cell assays (CAFCs), the trait mapped to a region on chromosome 18.¹⁸ Subsequently, variation in hematopoietic progenitor cell (HPC) frequency and HSC frequency were mapped in both young and old mice. This led to the identification of multiple QTLs, some of which were age- and differentiation stage-specific. Regardless of age, loci on chromosomes 7 and 18 were found to regulate HPC and HSC frequency, respectively. An additional locus on chromosome 1 was found to affect HPC and HSC frequency specifically in young mice, whereas loci on chromosomes 2 and 18 were found to affect these frequencies specifically in old mice.²⁷ A congenic mouse model was later used to confirm that the chromosome 2 locus indeed contained a regulator of HSC aging.²⁸ Variation in the percentage change of HSC frequency during aging was mapped to putative loci on chromosomes 2, 14 and X.²⁴

Particularly interesting was the finding that variation in both turnover rate of primitive hematopoietic cells and mouse lifespan mapped to overlapping regions on chromosomes 7 and 11.²⁹ This strengthened the hypothesis that mouse lifespan is in part dependent on progenitor turnover rate.

Yet another trait in which various inbred strains of mice have shown to differ is their absolute number of Lin⁻Sca1⁺⁺ cells and their responsiveness to early-acting cytokines, such as kit ligand, flt3 ligand and thrombopoietin. A genetic linkage study in BXD recombinant inbreds led to the identification of three loci

on chromosomes 2, 4 and 7 that affected the total number of Lin⁻Sca1⁺⁺ cells and a locus on chromosome 2 affecting their proliferative response to cytokines. The fact that both traits mapped to the exact same region on chromosome 2 suggests that the number of Lin⁻Sca1⁺⁺ cells may depend on their responsiveness to cytokines.³⁰ This hypothesis was reinforced when a QTL for the response of primitive cells to transforming growth factor- β 2 (TGF- β 2) was identified on chromosome 4 that overlapped with the previously identified QTL regulating the number of Lin⁻Sca1⁺⁺ cells.³¹

Furthermore, a strain-dependent variation was found to exist in the response to Granulocyte Colony-Stimulating factor (G-CSF); a growth factor that has the capacity to mobilize stem and progenitor cells from bone marrow into peripheral blood.³² When B6 (low responder), D2 (high responder), backcross and BXD mice were subjected to a genetic analysis, loci on chromosomes 2 and 11 (and possibly 4 and 14) were found to control G-CSF induced mobilization.³³

An overview of the QTLs that have been reported to associate with various hematopoietic traits is shown in Table 1. Interestingly, multiple traits have been mapped to the same QTL regions (note the “QTL-dense” regions on chromosomes 2 and 11), suggesting that they may be regulated by a common genetic element. It should also be noted that only a few QTLs have been independently replicated. In part, this is because relatively few laboratories have used quantitative trait genetics to study hematopoiesis. Secondly, it may be due to the inherently noisy quantitative nature of the traits under study as they are likely to be controlled by multiple QTLs that each have a limited effect. Finally, multiple *in vitro* and *in vivo* assays exist that measure the functional output of primitive hematopoietic cells. Although these assays are thought to have considerable overlap with each other, they may not be measuring exactly the same spectrum of cells. Thus, it is possible that these distinct cell subsets are controlled by different genetic elements.

The major limitation of classical QTL mapping approaches in recombinant inbred reference panels is that they are only able to identify genomic *regions* of interest, usually containing tens or even hundreds of genes. This poor resolution is due to the limited number of recombination events between the two sets of parental chromosomes. Of all the genes present within the QTL interval, it is hypothesized that only polymorphic variants can be responsible for variation in the observed phenotype. In most cases, this variability is in the form of single nucleotide polymorphisms (SNPs), and to a lesser extent in the form of deletions, insertions, rearrangements and copy number variations. Although every SNP has a potential impact on gene expression levels and therefore could affect HSC biology, the vast majority of SNPs are “synonymous” or “silent”.^{34;35} This makes it difficult to identify the causal variant or polymorphism, and therefore the causal gene that influences the trait of interest. An additional complication is the possibility that multiple linked (possibly even neighboring) genes collectively cause the phenotype, as was proposed for the progenitor cell cycling trait.³⁶

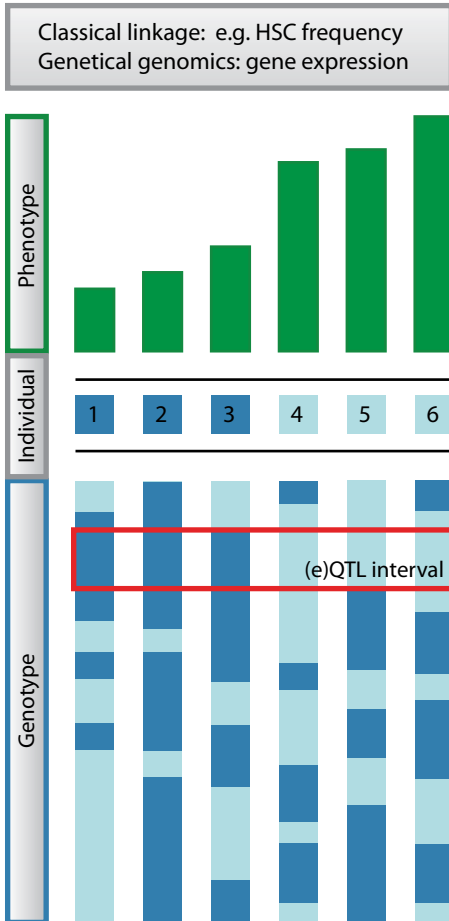


Figure 2. Overview of (expression) quantitative trait locus mapping procedure.

Variation in phenotype (here shown for six individuals) is correlated with variation in genotype (genotypes at a single chromosome are shown for each individual). The genomic location where these two parameters associate most strongly is referred to as the (expression) quantitative trait locus or (e)QTL. In this case, the three genetically distinct individuals that have a high value for the phenotype of interest carry the *light blue* genotype at the (e)QTL position, whereas the three that have a low phenotypic measure carry the *dark blue* genotype at that position. The phenotype can either be a classical trait (classical linkage) or the expression level of a gene (genetical genomics).

A promising approach that may aid in the identification of causal genes – and the networks in which they operate – is the combination of transcriptional profiling with linkage genetics.

Transcriptional profiling and linkage genetics combined

As explained, neither transcriptional profiling alone, nor genetic linkage analysis alone, has been shown to be an effective approach to identify genes or gene networks that specify “stemness”, initiate differentiation or govern lineage specification. However, the combination of both approaches may aid in their identification. Merging the fields of linkage genetics and genomics in this particular manner has been referred to as “genetical genomics”³⁷ or expression quantitative trait locus (eQTL) analysis.^{38,39} The genetical genomics approach considers individual gene expression levels to be quantitative traits. In cell

types isolated from genetically distinct individuals (e.g. BXD recombinant inbred mouse strains), linkage mapping can then be utilized to identify genomic regions affecting each gene expression trait (eQTL). The strategy of eQTL mapping is visualized in Figure 2. Hence, genetical genomics studies the genetic basis of variation in gene expression. When the genomic position of the gene and the eQTL which is associated with variation in its expression level coincide, the gene is considered to be cis-regulated. In contrast, when the eQTL associated with variation in its expression level maps to another position in the genome, the gene is considered to be trans-regulated. Cis-regulation is thought to arise from either local polymorphisms in the regulatory elements that alter gene expression levels, or alternatively to arise from polymorphisms in the coding region that affect mRNA stability or feedback regulation (Figure 3A). Cis-regulation can also originate from copy-number variability. Trans-regulation is thought to arise as a result of polymorphisms in the regulatory elements or coding region of a direct or indirect upstream regulator (Figure 3B). It should be emphasized that in this review the terms *cis*- and *trans*-regulation do not refer to the underlying molecular nature of the regulation, but only to the distance between the physical genomic position of a gene and its corresponding eQTL position. For this reason, it is possible that a gene can be classified as cis-regulated, even though it is actually regulated by one of its neighboring genes. To reduce the potential confusion between the type of regulation and the relative positions of genes and their eQTLs, the terms “local” and “distant” have also been proposed.⁴⁰

Brem *et al* were the first to report genetic mapping of global gene expression in a yeast cross.⁴¹ Since then, genetical genomics has been applied to genetically distinct strains of *Arabidopsis thaliana*, *Eucalyptus*, maize, *Caenorhabditis elegans*, mice, rats^{38,42-47} and also to cells isolated from human individuals.^{48,49} Collectively, these studies convincingly demonstrate the heritability of variation in transcript abundance and the presence of both cis- and trans-regulated genes. In addition, it is clear that the genetical genomics approach is broadly applicable to multiple species and cell types. A representative example of a genome-wide overview of genes and their eQTLs is depicted in Figure 3C. This eQTL regulator map was created by plotting the physical genomic position of variably expressed genes and the genomic positions that were most strongly associated with variation in their expression levels (eQTLs). Transcripts on the diagonal represent cis-regulated genes. Collections of transcripts that are identified as vertical bands or “transbands” represent genes that are located throughout the whole genome, but are thought to be transcriptionally affected by a common eQTL. If a certain genomic region harbors a higher frequency of eQTLs than expected by chance, it is termed an eQTL hotspot.^{38,41} The causal regulator within such an eQTL interval may be a signaling or transcription factor that affects the expression levels of its downstream targets.

Table 1. Quantitative trait loci (QTLs) associated with mouse hematopoietic traits.

Trait	Chr	Region (cM)	Reference ID	Ref. #
HPC frequency	1	38	Geiger et al. 2001	27
HSC frequency	1	38	Geiger et al. 2001	27
HSC frequency (LTC-IC)	1	52	Muller-Sieburg and Riblet 1996	17
Splenomegaly after G-CSF	1	53-75	Roberts et al. 2000	67
HSC frequency (LTC-IC)	1	99	Muller-Sieburg and Riblet 1996	17
HSC frequency	1	100	Geiger et al. 2001	27
Change in HSC frequency during aging	2	51	De Haan and Van Zant 1999	24
Lifespan	2	69	De Haan and Van Zant 1999	29
Progenitor cell mobilization by G-CSF	2	69?	Hasegawa et al. 2000	33
HPC frequency (old)	2	72	Geiger et al. 2001	27
HSC frequency (old)	2	72	Geiger et al. 2001	27
No. Lin ⁻ Sca1 ⁺ cells	2	73.2-77.2	Henckaerts et al. 2002	30
Response of primitive cells to cytokines	2	73.2-77.2	Henckaerts et al. 2002	30
HSC frequency	3	24	Geiger et al. 2001	27
Lifespan	4	40.3	De Haan and Van Zant 1999	29
Response of primitive cells to TGF- β 2	4	67-79	Langer et al. 2004	31
No. Lin ⁻ Sca1 ⁺ cells	4	77.5-79	Henckaerts et al. 2002	30
Progenitor cell cycling	4	78	De Haan and Van Zant 1999	29
Progenitor cell mobilization by G-CSF	4	?	Hasegawa et al. 2000	33
HSC frequency	5	20	Geiger et al. 2001	27
Progenitor cell cycling	7	0.5	De Haan et al. 2002	36
Progenitor cell cycling	7	6	De Haan and Van Zant 1999	29
Lifespan	7	6	De Haan and Van Zant 1999	29
HPC frequency	7	8	Geiger et al. 2001	27
HPC frequency	7	25	Geiger et al. 2001	27
HPC frequency	7	50	Geiger et al. 2001	27
HPC frequency (old)	7	50	Geiger et al. 2001	27
No. Lin ⁻ Sca1 ⁺ cells	7	51-53	Henckaerts et al. 2002	30
Progenitor cell cycling	9	61	De Haan and Van Zant 1999	29
Progenitor cell mobilization by G-CSF	11	3	Hasegawa et al. 2000	33
Progenitor cell cycling	11	29	De Haan et al. 2002	36
Progenitor cell cycling	11	31	De Haan and Van Zant 1999	29
Lifespan	11	31	De Haan and Van Zant 1999	29
HSC frequency (LTC-IC)	11	32	Muller-Sieburg and Riblet 1996	17
Change in HSC frequency during aging	14	12	De Haan and Van Zant 1999	24
HPC frequency (old)	14	13	Geiger et al. 2001	27
Progenitor cell mobilization by G-CSF	14	?	Hasegawa et al. 2000	33
HSC frequency	15	43	De Haan and Van Zant 1997	18
HPC frequency (old)	15	59	Geiger et al. 2001	27
HSC frequency	18	19	De Haan and Van Zant 1997	18
HSC frequency	18	27	Geiger et al. 2001	27
HPC frequency (old)	18	27	Geiger et al. 2001	27
HSC frequency (old)	18	27	Geiger et al. 2001	27
Change in HSC frequency during aging	X	5	De Haan and Van Zant 1999	24
HSC frequency	X	51	Geiger et al. 2001	27
Response of primitive cells to cytokines	X	?	Henckaerts et al. 2002	30

All traits were studied in C57Bl/6, DBA/2, backcross and/ or BXD recombinant inbred mice. Unless noted otherwise, traits were analyzed in young mice, and HPC and HSC frequencies were measured using CAFC assays. Italic font indicates traits and corresponding QTLs that are mentioned elsewhere in this review. Boxes indicate "QTL-dense" regions. (Note that not all of the reported genetic associations have met the most stringent statistical threshold for significant genome-wide linkage. Also note that in the past decade, the mouse genome map has undergone significant revisions, and therefore the precise genomic locations of the identified QTLs may be slightly inaccurate, especially in older publications. Nevertheless, for historical accuracy, this table shows QTL regions as specified in the original references.)

Abbreviations: CAFC, cobblestone area forming cell; G-CSF, granulocyte colony-stimulating factor; HPC, hematopoietic progenitor cell; HSC, hematopoietic stem cell; LTC-IC, long-term culture-initiating cell; TGF, transforming growth factor; ?, exact region not specified.

Associated with each eQTL are two important sets of genes: those regulated by the eQTL (transband genes if more than one) and those within the eQTL interval that are candidate regulators. These regulators can either directly or indirectly affect the abundance of the transband transcripts, raising the prospect of several levels of hierarchy within the regulatory network (Figure 3D).

To identify the causal gene within each eQTL, the interval is first screened for genetic variability, since only those genes that are polymorphic can have a functional impact on the transcript abundance of the transband genes. Both regulatory elements and coding regions must be analyzed for the presence of polymorphisms. In particular, cis-acting genes are high-priority candidate regulators, as they may contain genetic variants that not only influence their own expression levels, but also those of the transband genes. Another category of genes which could affect the expression of the transband genes are those that are equally expressed, but carry polymorphisms in their coding regions, resulting in the generation of functionally distinct proteins with a differential ability to regulate downstream targets. For example, these polymorphisms can alter protein functionality by being “non-synonymous” (amino acid changing) or alternatively by introducing alternative splicing. It should be noted that this category of candidate regulators would remain unnoticed in traditional micro-array experiments, since the expression levels of such transcripts would not necessarily be altered.

While it is tempting to assume that trans-regulated genes preferentially map to eQTL intervals containing transcription factors, in yeast this could not be verified.⁵⁰ Another potential category of regulatory genes consists of signaling factors that can indirectly affect the abundance of the transband transcripts.

Although the terms cis- and trans-regulation are commonly accepted descriptions of gene regulatory relationships, their biological relevance is only assumed. While it is clear that genetical genomics has great potential to identify novel regulatory pathways and increase our understanding of regulatory networks, functional validation of candidate regulators is ultimately necessary to confirm their biological activity.

HSCs and genetical genomics

Classical QTL mapping results in the identification of genomic intervals that affect traits of interest. This approach narrows down the number of candidate genes affecting such traits from around 30,000 (all genes in the genome) to tens or hundreds of genes (those located within that QTL interval). However, despite this major improvement, for a molecular biologist it would still require a huge effort to functionally test all the candidate genes in that interval.

A complementary approach to assist in the identification of candidate genes affecting stem cell traits is genetical genomics. Using the Affymetrix gene expression platform we collected data for each of the inbred strains from the BXD reference panel and the variation in transcript abundance in primitive

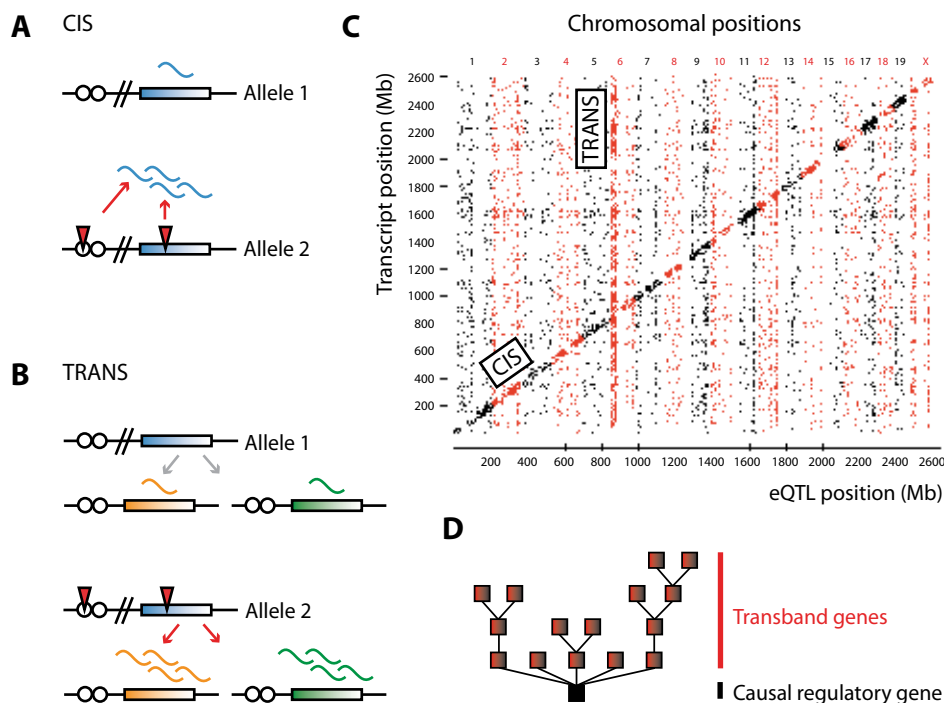


Figure 3. Cis- and trans-regulated gene expression. (A) Cis-regulation is expected to originate from polymorphisms (red triangles) in the regulatory elements (white circles) or the coding region (colored rectangle) of the gene itself (or possibly of a nearby gene). (B) Trans-regulation is expected to originate from polymorphisms in the regulatory elements or the coding region of a gene located distant from the gene whose expression it controls. Note that variation in expression of multiple genes can map to the same gene in trans. (C) Transcripts and their eQTLs are graphically depicted in a genome-wide eQTL regulator map. Plotted on the y-axis are the physical positions of all measured transcripts, whereas on the x-axis the genomic regions that are most strongly associated with variation in expression levels (i.e. eQTLs) of the corresponding transcripts are shown. When transcript and eQTL position coincide, the transcript is considered to be cis-regulated and plotted on the diagonal. The vertical transband refers to transcripts encoded by genes that are positioned throughout the whole genome, but map to the same eQTL position. Transband transcripts are suggested to be co-regulated. Potential transband regulators are located within the eQTL interval (where the transband meets the x-axis). Figure adapted from ⁴². (D) Co-regulated transband genes can be directly or indirectly targeted by the potential regulator, thereby creating a network that consists of multiple levels of gene regulation.

Lin-Sca1⁺cKit⁺ cells of each gene was mapped to an eQTL. Using stringent significance thresholds, a total of 162 cis-regulated and 136 trans-regulated genes were identified. *Runx1*, a well-known hematopoietic regulator,^{51;52} was found to be strongly cis-regulated. Interestingly, two of its known downstream targets (*Tcrb* and *Csf1r*) were found to be co-regulated with *Runx1*, indicating the biological relevance of the networks that could be identified. The generation of this comprehensive eQTL dataset allowed the more detailed analysis of QTL

intervals that had previously been identified using the classical approach. Cis-regulated genes within these intervals represent the best candidate regulators, as they may harbor genetic variants that affect both their own expression levels and the trait of interest. Within the QTL interval on chromosome 11 to which we had previously mapped a region associated with hematopoietic progenitor cell turnover, only eight cis-regulated genes were identified, thereby reducing the number of candidate genes for functional testing.⁴²

An example of the power of this combined approach was recently provided by Liang *et al.* Classical QTL analysis had previously identified regions on chromosomes 3, 5 and 18 that associated with variation in HSC frequency.²⁷ Using reciprocal congenic mouse strains, the chromosome 3 QTL interval in isolation was subsequently shown to be sufficient to confer this stem cell phenotype. Thereafter, a detailed analysis of differentially expressed transcripts within the QTL interval, followed by functional confirmation, led to the identification of *Lxn* as a gene involved in determining HSC frequency.⁵³

It should be noted that large collections of expression data have been deposited in the online database GeneNetwork (<http://www.genenetwork.org>) and are freely accessible to the research community.⁵⁴ GeneNetwork contains genotypic, phenotypic and gene expression data from several species, including *Arabidopsis thaliana*, barley, mouse and rat. Within the mouse BXD reference population, expression data of multiple tissues (HSCs, regulatory T cells, various neural tissues, eye, liver, lung and kidney) are present, which permit the distinction between genes that are expressed or regulated in a tissue-specific manner from those that are equally expressed or regulated in multiple tissues.

Multi-dimensional genetical genomics

Genetical genomics has proven to be a valuable tool for the identification of genes and gene networks that operate in HSCs. Yet, its potential impact is only emerging and has not been fully exploited. Therefore, in this section we will address the future applications of genetical genomics.

Adding the dimension of closely related cell types

Genetical genomics has primarily been limited to single cell types. A more powerful approach would be to apply the same approach to closely related cell types. This would allow comparative analyses of gene regulatory networks between distinct but related cells. In the HSC field such multi-dimensional genetical genomics studies have not yet been performed. However, Li *et al* recently demonstrated the general validity of the approach through the application of genetical genomics to *Caenorhabditis elegans* recombinant inbred strains that were exposed to different temperatures. Their results showed heritable variation in gene expression responses to these environmental changes.⁴⁷ This has created a solid basis for future multi-dimensional genetical genomics approaches.

In the hematopoietic system, genetical genomics could conceivably be applied to different hematopoietic cell stages (e.g. primitive HSCs, committed progenitors and fully differentiated blood cells). Inclusion of this additional dimension would enable the study of cell fate decisions during the process of hematopoietic cell differentiation. Whereas classical analysis of gene expression levels during differentiation evaluates the dynamics of gene *expression*, analyzing eQTLs during differentiation evaluates the dynamics of gene *regulation*, permitting the identification of genes and gene networks that are specifically active in one cell type and not in another. Although eQTL profiles of different species and different cell types within the same species have previously been compared, eQTL profiles have never been evaluated for highly purified cell types that are so closely related. By implementing a “subtractive genetical genomics approach” a distinction can possibly be made between common or “housekeeping” eQTLs and those eQTLs that are specific for only one cell stage.

Multi-dimensional genetical genomics would also be a powerful tool to study age-dependent changes in the HSC compartment. Previous studies have compared the transcriptional profiles of HSC-enriched populations from young and old murine bone marrow. The collections of genes reported to be up-regulated in aged stem cells included those involved in inflammatory and stress responses⁵⁵ and signal transducer activity and receptor activity,⁵⁶ whereas those down-regulated during aging were genes involved in DNA repair and chromatin remodelling.⁵⁵ While these studies revealed thousands of age-regulated genes, the ultimate causes of these expression perturbations remain unknown. Analyzing age-dependent gene expression changes using multi-dimensional genetical genomics could bring the identification of genes causing the age-induced alterations – and thereby future therapeutic intervention strategies – one step closer.

Adding the dimension of epigenetics

Epigenetic gene regulation has been suggested to play a key role in modulating stem cell fate. Epigenetics refers to heritable gene expression changes that occur without DNA sequence alterations, and includes DNA methylation and histone modifications such as acetylation, methylation, and ubiquitylation. These modifications can result in either gene activation or gene repression. That epigenetic gene repression appears to be involved in the maintenance of “stemness” became apparent when a number of developmental regulators were found to be epigenetically silenced in murine ESCs and activated upon induction of ESC differentiation.^{57;58} It is exciting to postulate that epigenetic modifiers might similarly define cellular fate and lineage commitment during hematopoiesis.

Since a differential epigenetic conformation of the genome can result in variation in gene expression levels that can in turn affect stem cell traits, it is possible that such differential epigenetic states underlie some of the observed (e)QTL effects. At this time, however, there is insufficient knowledge on how

epigenetic modifications correlate with variation in gene expression levels on a genome-wide scale. To this end, a powerful approach would be to combine traditional transcriptome profiling with whole-genome tiling arrays measuring chromatin-immunoprecipitation and DNA methylation. If these complementary array-based analyses were carried out in the same reference panel of genetically distinct individuals, not only variation in transcript abundance, but potentially also variation in epigenetic conformations could be mapped to genomic loci. In this manner, both the genetics of gene expression (i.e. “regular” genetical genomics) and the genetics of epigenetics could be studied simultaneously, thus revealing genes that directly or indirectly affect epigenetic gene states. An additional issue that could be addressed by such an approach is to estimate the percentage of variation in gene expression that can be explained by different epigenetic conformations.

The level of complexity could be further increased by including different cell types in the analysis, such as the above-mentioned different hematopoietic cell stages, different stem cell types, stem cells derived from different species, or stem cells of different ages. Through a “subtractive QTL analysis” approach, cell type specific QTLs that affect transcript abundance and/ or epigenetic gene states might thus be identified. The emergence of such comparative analyses in the coming years will further the understanding of regulatory networks and how they affect cellular fate.

Adding the dimension of microRNAs

It is possible that some of the upstream regulators located within (e)QTL intervals are in fact not protein-encoding genes, but rather microRNAs. MicroRNAs are small non-coding RNAs complementary to one or often multiple mRNAs, and their main function appears to be down-regulation of gene expression. Certain microRNAs have been shown to be differentially expressed between various hematopoietic cell types, suggesting that they could be involved in lineage specification,⁵⁹ Polymorphisms in microRNA production sites can alter their specificity, whereas polymorphisms in regulatory elements can alter their expression levels. In addition, polymorphisms in microRNA target sites can affect the binding of a microRNA and therefore their capability to silence target gene expression. Any of these microRNA sequence variants might underlie (e) QTL effects. For example, polymorphic microRNA target sites located within a gene (often in the 3' untranslated region) might underlie a cis-acting eQTL effect. Further, transbands could be explained by a polymorphic microRNA production site that is positioned within an eQTL interval, giving rise to a microRNA that may differentially affect the expression levels of its target genes.

Efforts have already been made to overlap polymorphic microRNA target sites with eQTL intervals and known classical QTL intervals in an attempt to identify microRNAs that not only underlie variation in gene expression levels but also variation in cell biological traits. Genes positioned in (e)QTL intervals can be

screened for polymorphisms in microRNA target sites in an online database at <http://compbio.utmem.edu/miRSNP/>.⁶⁰

Little is known about the factors that regulate the expression levels of the microRNAs themselves. Genome-wide microRNA profiling of cells isolated from genetically distinct individuals would assist in this regard, since through the use of genetic linkage, variation in microRNA expression levels can be mapped to genomic regions affecting microRNA expression levels.

Adding the dimension of clinical data

The sequencing of the human genome and the development of transcriptome profiling technology have permitted new approaches to characterize hematologic malignancies at the molecular level. Gene expression profiles have been generated for malignancies such as diffuse large B-cell lymphoma, mantle cell lymphoma, acute myeloid leukemia, acute lymphoblastic leukemia, chronic lymphocytic leukemia and multiple myeloma. All these malignancies could be classified into molecularly distinct subgroups on the basis of similarities in their gene expression profiles, and genes whose expression could discriminate between these distinct subgroups were identified (reviewed in ⁶¹). Although this subgroup classification has diagnostic, prognostic and therapeutic consequences, the disease-initiating or causative factors are still not known. If large scale clinical data and gene expression profiles were combined with detailed genotypes of the patients, this would permit the use of genetical genomics and therefore the identification of QTLs that underlie complex diseases, and contribute to understanding which genes, gene networks and biological processes are involved in both normal and malignant hematopoietic cell development.

That genetical genomics could be applicable to human data was demonstrated in two independent studies using previously genotyped lymphoblastoid cell lines from related individuals.^{48,49} In both studies the heritability of gene expression levels was shown and eQTLs were identified, but there were too many differences between the two approaches to compare them in a direct manner.⁶² Association-based studies were also performed using lymphoblastoid cell lines from unrelated individuals.^{63,64} While these studies demonstrated the potential of performing genetical genomics using human data, the clinical relevance of these approaches was limited since the studies were performed on transformed human cell lines, were limited in their samples sizes, and did not include any clinical phenotypes.

A more direct example of the clinical relevance of this approach was recently provided by Göring *et al*, who generated genome-wide transcriptional profiles of normal untransformed lymphocytes from a large collection of genotyped individuals whose plasma cholesterol concentrations were also measured. Using a genetical genomics approach, *VNN1* was identified as a gene affecting high-density lipoprotein cholesterol concentrations.⁶⁵

In a second clinically relevant example, blood and adipose tissues were collected from a large group of subjects, after which genotypes, gene expression

levels and clinical traits related to obesity were analyzed in a combinatorial fashion. A substantial correlation was found between gene expression profiles of adipose tissue and obesity-related traits, but not between blood expression profiles and those traits. Importantly, genes and gene networks that were enriched in inflammatory and immune response pathways were identified that in part contribute to obesity in humans.⁶⁶ These examples demonstrate how a multi-dimensional genetical genomics approach can aid in the understanding of human health.

CLOSING REMARKS

In this review, we describe how transcriptional profiling has helped to define the molecular identity of HSCs and other cell types, and how the use of linkage genetics has permitted the identification of specific genomic regions that affect HSC traits. In particular, we focus on the “genetical genomics” approach of combining transcriptional profiling with genetic linkage analysis, and discuss the potential added value of including additional dimensions in the analysis. All approaches are summarized in Figure 4. In the coming years, multi-dimensional genetical genomics has the potential to greatly aid in revealing regulatory networks that specify cell fate decisions not only in HSCs, but in a whole range of clinically relevant cell types.


Transcriptome profiling	<i>One genotype</i> Within one or more cell types	Gene expression
Classical linkage	<i>Multiple genotypes</i> Within organism	Classical phenotype
One-dimensional genetical genomics	<i>Multiple genotypes</i> Within one cell type	Gene expression levels
Multi-dimensional genetical genomics 	<i>Multiple genotypes</i> Within multiple cell types: Developmental stages Unstimulated/ stimulated Healthy/ diseased Young/ old Etc.	Gene expression levels DNA methylation status Histone modification status Protein levels Protein phosphorylation levels Etc.

Figure 4. Overview of the described approaches (left), the sources of analysis (middle) and the phenotypic measures (right).

ACKNOWLEDGEMENTS

2

This work was supported by the Netherlands Genomics Initiative (Horizon, 050-71-055); the Netherlands Organization for Scientific Research (VENI, 916-86-009 to B.D.; and VICI, 918-76-601 to G.d.H.); and by the European Community (Marie Curie RTN EUrythron, MRTN-CT-2004-005499).

REFERENCES

1. Weissman IL, Anderson DJ, Gage F. Stem and progenitor cells: origins, phenotypes, lineage commitments, and transdifferentiations. *Annu.Rev.Cell Dev.Biol.* 2001;17:387-403.
2. Ivanova NB, Dimos JT, Schaniel C et al. A stem cell molecular signature. *Science* 2002;298(5593):601-604.
3. Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA. "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* 2002;298(5593):597-600.
4. Fortunel NO, Otu HH, Ng HH et al. Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science* 2003;302(5644):393.
5. Burns CE, Zon LI. Portrait of a stem cell. *Dev.Cell* 2002;3(5):612-613.
6. Evsikov AV, Solter D. Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science* 2003;302(5644):393.
7. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006;126(4):663-676.
8. Wernig M, Meissner A, Cassady JP, Jaenisch R. C-Myc is dispensable for direct reprogramming of mouse fibroblasts. *Cell Stem Cell* 2008;2(1):10-12.
9. Nakagawa M, Koyanagi M, Tanabe K et al. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat.Biotechnol.* 2008;26(1):101-106.
10. Hanna J, Markoulaki S, Schorderet P et al. Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell* 2008;133(2):250-264.
11. Yu J, Vodyanik MA, Smuga-Otto K et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science* 2007;318(5858):1917-1920.
12. Park IH, Zhao R, West JA et al. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* 2008;451(7175):141-146.
13. Takahashi K, Tanabe K, Ohnuki M et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 2007;131(5):861-872.
14. Enver T, Greaves M. Loops, lineage, and leukemia. *Cell* 1998;94(1):9-12.
15. Chambers SM, Boles NC, Lin KYK et al. Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell* 2007;1578-591.
16. De Haan G, Nijhof W, Van Zant G. Mouse strain-dependent changes in frequency and proliferation of hematopoietic stem cells during aging: correlation between lifespan and cycling activity. *Blood* 1997;89(5):1543-1550.
17. Muller-Sieburg CE, Riblet R. Genetic control of the frequency of hematopoietic stem cells in mice: mapping of a candidate locus to chromosome 1. *J.Exp.Med.* 1996;183(3):1141-1150.
18. De Haan G, Van Zant G. Intrinsic and extrinsic control of hemopoietic stem cell numbers: mapping of a stem cell gene. *J.Exp.Med.* 1997;186(4):529-536.
19. Van Zant G, Eldridge PW, Behringer RR, Dewey MJ. Genetic control of hematopoietic kinetics revealed by analyses of allophenic mice and stem cell suicide. *Cell* 1983;35(3 Pt 2):639-645.
20. Harrison DE, Astle CM, Stone M. Numbers and functions of transplantable primitive immunohematopoietic stem cells. Effects of age. *J.Immunol.* 1989;142(11):3833-3840.
21. Morrison SJ, Wandycz AM, Akashi K, Globerson A, Weissman IL. The aging of hematopoietic stem cells. *Nat.Med.* 1996;2(9):1011-1016.
22. Sudo K, Ema H, Morita Y, Nakauchi H. Age-associated characteristics of murine

- hematopoietic stem cells. *J.Exp.Med.* 2000;192(9):1273-1280.
23. Liang Y, Van Zant G, Szilvassy SJ. Effects of aging on the homing and engraftment of murine hematopoietic stem and progenitor cells. *Blood* 2005;106(4):1479-1487.
 24. De Haan G, Van Zant G. Dynamic changes in mouse hematopoietic stem cell numbers during aging. *Blood* 1999;93(10):3294-3301.
 25. Chen J, Astle CM, Harrison DE. Genetic regulation of primitive hematopoietic stem cell senescence. *Exp.Hematol.* 2000;28(4):442-450.
 26. Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC.Genet.* 2004;57.
 27. Geiger H, True JM, De Haan G, Van Zant G. Age- and stage-specific regulation patterns in the hematopoietic stem cell hierarchy. *Blood* 2001;98(10):2966-2972.
 28. Geiger H, Rennebeck G, Van Zant G. Regulation of hematopoietic stem cell aging in vivo by a distinct genetic element. *Proc.Natl.Acad.Sci.U.S.A* 2005;102(14):5102-5107.
 29. De Haan G, Van Zant G. Genetic analysis of hemopoietic cell cycling in mice suggests its involvement in organismal life span. *FASEB J.* 1999;13(6):707-713.
 30. Henckaerts E, Geiger H, Langer JC et al. Genetically determined variation in the number of phenotypically defined hematopoietic progenitor and stem cells and in their response to early-acting cytokines. *Blood* 2002;99(11):3947-3954.
 31. Langer JC, Henckaerts E, Orenstein J, Snoeck HW. Quantitative trait analysis reveals transforming growth factor-beta2 as a positive regulator of early hematopoietic progenitor and stem cell function. *J.Exp.Med.* 2004;199(1):5-14.
 32. Roberts AW, Foote S, Alexander WS et al. Genetic influences determining progenitor cell mobilization and leukocytosis induced by granulocyte colony-stimulating factor. *Blood* 1997;89(8):2736-2744.
 33. Hasegawa M, Baldwin TM, Metcalf D, Foote SJ. Progenitor cell mobilization by granulocyte colony-stimulating factor controlled by loci on chromosomes 2 and 11. *Blood* 2000;95(5):1872-1874.
 34. Kimura M. Evolutionary rate at the molecular level. *Nature* 1968;217(5129):624-626.
 35. King JL, Jukes TH. Non-Darwinian evolution. *Science* 1969;164(881):788-798.
 36. De Haan G, Bystrykh LV, Weersing E et al. A genetic and genomic analysis identifies a cluster of genes associated with hematopoietic cell turnover. *Blood* 2002;100(6):2056-2062.
 37. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Genet.* 2001;17(7):388-391.
 38. Schadt EE, Monks SA, Drake TA et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 2003;422(6929):297-302.
 39. Schadt EE, Lamb J, Yang X et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat.Genet.* 2005;37(7):710-717.
 40. Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat.Rev.Genet.* 2006;7(11):862-872.
 41. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science* 2002;296(5568):752-755.
 42. Bystrykh L, Weersing E, Dontje B et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat.Genet.* 2005;37(3):225-232.
 43. Chesler EJ, Lu L, Shou S et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* 2005;37(3):233-242.
 44. DeCook R, Lall S, Nettleton D, Howell SH. Genetic regulation of gene expression during shoot development in Arabidopsis. *Genetics* 2006;172(2):1155-1164.
 45. Hubner N, Wallace CA, Zimdahl H et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat.Genet.* 2005;37(3):243-253.
 46. Kirst M, Myburg AA, De Leon JP et al. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. *Plant Physiol* 2004;135(4):2368-2378.
 47. Li Y, Alvarez OA, Gutteling EW et al. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS.Genet.* 2006;2(12):e222.
 48. Monks SA, Leonardson A, Zhu H et al. Genetic inheritance of gene expression in human cell lines. *Am.J.Hum.Genet.* 2004;75(6):1094-1105.
 49. Morley M, Molony CM, Weber TM et al. Genetic analysis of genome-wide

- variation in human gene expression. *Nature* 2004;430(7001):743-747.
50. Yvert G, Brem RB, Whittle J et al. Transacting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat.Genet.* 2003;35(1):57-64.
 51. Okuda T, Van Deursen J, Hiebert SW, Grosveld G, Downing JR. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell* 1996;84(2):321-330.
 52. Wang Q, Stacy T, Binder M et al. Disruption of the *Cbfa2* gene causes necrosis and hemorrhaging in the central nervous system and blocks definitive hematopoiesis. *Proc.Natl.Acad.Sci.U.S.A* 1996;93(8):3444-3449.
 53. Liang Y, Jansen M, Aronow B, Geiger H, Van Zant G. The quantitative trait gene *latexin* influences the size of the hematopoietic stem cell population in mice. *Nat.Genet.* 2007;39(2):178-188.
 54. Chesler EJ, Lu L, Wang J, Williams RW, Manly KF. WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nat. Neurosci.* 2004;7(5):485-486.
 55. Chambers SM, Shaw CA, Gatz C et al. Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation. *PLoS.Biol.* 2007;5(8):e201.
 56. Rossi DJ, Bryder D, Zahn JM et al. Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proc.Natl.Acad.Sci.U.S.A* 2005;102(26):9194-9199.
 57. Boyer LA, Plath K, Zeitlinger J et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 2006;441(7091):349-353.
 58. Bernstein BE, Mikkelsen TS, Xie X et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 2006;125(2):315-326.
 59. Chen CZ, Li L, Lodish HF, Bartel DP. MicroRNAs modulate hematopoietic lineage differentiation. *Science* 2004;303(5654):83-86.
 60. Bao L, Zhou M, Wu L et al. PolymiRST Database: linking polymorphisms in microRNA target sites with complex traits. *Nucleic Acids Res.* 2007;35(Database issue):D51-D54.
 61. Margalit O, Somech R, Amariglio N, Rechavi G. Microarray-based gene expression profiling of hematologic malignancies: basic concepts and clinical applications. *Blood Rev.* 2005;19(4):223-234.
 62. De Koning DJ, Haley CS. Genetical genomics in humans and model organisms. *Trends Genet.* 2005;21(7):377-381.
 63. Stranger BE, Forrest MS, Clark AG et al. Genome-wide associations of gene expression variation in humans. *PLoS. Genet.* 2005;1(6):e78.
 64. Cheung VG, Spielman RS, Ewens KG et al. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 2005;437(7063):1365-1369.
 65. Goring HH, Curran JE, Johnson MP et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat.Genet.* 2007;39(10):1208-1216.
 66. Emilsson V, Thorleifsson G, Zhang B et al. Genetics of gene expression and its effect on disease. *Nature* 2008;452(7186):423-428.
 67. Roberts AW, Hasegawa M, Metcalf D, Foote SJ. Identification of a genetic locus modulating splenomegaly induced by granulocyte colony-stimulating factor in mice. *Leukemia* 2000;14(4):657-661.

CHAPTER

3

EXPRESSION QUANTITATIVE TRAIT LOCI ARE HIGHLY SENSITIVE TO CELLULAR DIFFERENTIATION STATE

Alice Gerrits,* Yang Li,* Bruno M. Tesson,*
Leonid V. Bystrykh, Ellen Weersing, Albertina Ausema,
Bert Dontje, Xusheng Wang, Rainer Breitling,
Ritsert C. Jansen and Gerald de Haan

** These authors contributed equally to this work*

*PLoS Genetics. 2009 Oct;5(10):e1000692
Highlight in Nature Rev Genet. 2009 10 (12):819*

ABSTRACT

Genetical genomics is a strategy for mapping gene expression variation to expression quantitative trait loci (eQTLs). We performed a genetical genomics experiment in four functionally distinct but developmentally closely related hematopoietic cell populations isolated from the BXD panel of recombinant inbred mouse strains. This analysis allowed us to analyze eQTL robustness/sensitivity across different cellular differentiation states. Although we identified a large number (365) of “static” eQTLs that were consistently active in all four cell types, we found a much larger number (1283) of “dynamic” eQTLs showing cell-type-dependence. Of these, 140, 45, 531, and 295 were preferentially active in stem, progenitor, erythroid and myeloid cells, respectively. A detailed investigation of those *dynamic* eQTLs showed that in many cases the eQTL specificity was associated with expression changes in the target gene. We found no evidence for target genes that were regulated by distinct eQTLs in different cell types, suggesting that large-scale changes within functional regulatory networks are uncommon. Our results demonstrate that heritable differences in gene expression are highly sensitive to the developmental stage of the cell population under study. Therefore, future genetical genomics studies should aim at studying multiple well-defined and highly purified cell types in order to construct as comprehensive a picture of the changing functional regulatory relationships as possible.

AUTHOR SUMMARY

Blood cell development from multipotent hematopoietic stem cells to specialized blood cells is accompanied by drastic changes in gene expression for which the triggers remain mostly unknown. Genetical genomics is an approach linking natural genetic variation to gene expression variation, thereby allowing the identification of genomic loci containing gene expression modulators (eQTLs). In this paper, we used a genetical genomics approach to analyze gene expression across four developmentally close blood cell types collected from a large number of genetically different but related mouse strains. We found that while a significant number of eQTLs (365) had a consistent “static” regulatory effect on gene expression, an even larger number were found to be very sensitive to cell stage. As many as 1283 eQTLs exhibited a “dynamic” behavior across cell types. By looking more closely at these *dynamic* eQTLs, we show that the sensitivity of eQTLs to cell stage is largely associated with gene expression changes in target genes. These results stress the importance of studying gene expression variation in well-defined cell populations. Only such studies will be able to reveal the important differences in gene regulation between different cell types.

INTRODUCTION

Genetical genomics uses quantitative genetics on a panel of densely genotyped individuals to map genomic loci that modulate gene expression.¹ The quantitative trait loci identified in this manner are referred to as expression quantitative trait loci, or eQTLs.² Most genetical genomics studies that have thus far been reported have analyzed single cell types or compared developmentally unrelated and distant cell types.³⁻⁸ Here, we report the first application of genetical genomics to study eQTL dynamics across closely related cell types during cellular development. We show results that discriminate between eQTLs that are consistently active or “static” and those that are cell-type-dependent or “dynamic”.

We used the hematopoietic system as a model to analyze how the genome of a single stem cell is able to generate a large variety of morphologically and functionally distinct differentiated cells. Differentiation of hematopoietic stem cells towards mature, lineage-committed blood cells is associated with profound changes in gene expression patterns. The search for differentially expressed genes, most notably for those transcripts exclusively present in stem cells and not in their more differentiated offspring, has been successful and has provided valuable insight into the molecular nature of stem cell self-renewal.⁹⁻¹² Yet, complementary approaches were needed to elucidate the dynamic regulatory pathways that are underlying the robust differentiation program leading to blood cell production.

We describe a genetic analysis of variation in gene expression across four functionally distinct, but developmentally related hematopoietic cell populations. Our data reveal complex cell-stage specific patterns of heritable variation in transcript abundance, demonstrating the plasticity of gene regulation during hematopoietic cell differentiation.

METHODS

Recombinant inbred mice

Female BXD recombinant inbred mice were originally purchased from The Jackson Laboratory and housed under clean conventional conditions. Mice were used between 3 and 4 months of age. All animal experiments were approved by the Groningen University Animal Care Committee.

Cell purification

Bone marrow cells were flushed from the femurs and tibias of three mice and pooled. After standard erythrocyte lysis, nucleated cells were stained with either a panel of biotin-conjugated lineage-specific antibodies (containing antibodies to CD3e, CD11b (Mac1), CD45R/ B220, Gr-1 (Ly-6G and Ly-6C) and TER-119 (Ly-76)), fluorescein isothiocyanate (FITC)-conjugated antibody to Sca-1 and allophycocyanin (APC)-conjugated antibody to c-Kit, or with biotin-conjugated TER-119 antibody and FITC-conjugated antibody to Gr-1. After being washed, cells were incubated with streptavidin-phycoerythrin (PE) (all antibodies were purchased from Pharmingen). Cells were purified using a MoFlo flowcytometer (BeckmanCoulter) and were immediately collected in RNA lysis buffer. Lineage-depleted (Lin^-) bone marrow cells were defined as the 5% of cells showing the least PE intensity.

RNA isolation and Illumina microarrays

Total RNA was isolated using the RNeasy Mini kit (Qiagen) in accordance with the manufacturer's protocol. RNA concentration was measured using a Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies). The RNA quality and integrity was determined using Lab-on-Chip analysis on an Agilent 2100 Bioanalyzer (Agilent Technologies). Biotinylated cRNA was prepared using the Illumina TotalPrep RNA Amplification Kit (Ambion) according to the manufacturer's specifications starting with 100 ng total RNA. Per sample, 1.5 μg of cRNA was used to hybridize to Sentrix Mouse-6 BeadChips (Illumina). Hybridization and washing were performed by ServiceXS according to the Illumina standard assay procedures. Scanning was carried out on the Illumina BeadStation 500. Image analysis and extraction of raw expression data were performed with Illumina Beadstudio v2.3 Gene Expression software with default settings and no normalization. The raw

expression data from all four cell types were first log₂ transformed and then quantile normalized as a single group.

Clustering of genes

For cluster analysis we retained only genes having a minimal fold change of 2 (difference of 1 in log₂ scale) in either direction in mean expression on the transition from Lin⁻Sca-1⁺c-Kit⁺ to Lin⁻Sca-1⁻c-Kit⁺ and on the transition from Lin⁻Sca-1⁻c-Kit⁺ to TER-119⁺ or to Gr-1⁺. This filter reduced the dataset to 876 probes. We then computed the distance matrix for this group of probes, using the absolute Pearson correlation. Using this distance matrix, we applied the hierarchical clustering algorithm. From the resulting tree, 8 different clusters emerged from a manually chosen threshold. We then submitted each of these clusters to DAVID to identify enriched functional annotations.¹³

Full ANOVA model for eQTL mapping

The expression data of the four cell types were firstly corrected for batch effect and then analyzed separately by the following ANOVA model:

$$y_i = \mu + Q_i + e_i$$

where y_i is the gene's log intensity on the i th microarray; μ is the mean; Q_i is the genotype effect under study; and e_i is the residual error.

Next, expression data of the four cell types were combined and analyzed by a full ANOVA model including the cell type effect (CT) and the eQTL×CT interaction effect:

$$y_{ij} = \mu + CT_j + Q_i + (Q \times CT)_{ij} + e_{ij}$$

where y_{ij} is the gene's log intensity at the i th microarray ($i = 1, \dots, n$) and j th cell type; CT_j is the j th cell type effect; $(Q \times CT)_{ij}$ is the interaction effect between the i th eQTL genotype and j th cell type, and e_{ij} is the residual error. The batch effect was included as one of the factors. For each probe, we performed a genome-wide linkage analysis to identify the two markers that showed the most significant main QTL effect and interaction effect, respectively.

Local and distant eQTLs

We defined an eQTL as *local* if it was located within less than 10 Mb from the gene. All other eQTLs were considered *distant*.

Classification of eQTLs

The ANOVA yields significance p -values for the main QTL effect Q_i and the interaction effect $(Q \times CT)_{ij}$ for each probe at each marker. A small p -value for the interaction effect indicates that the eQTL effect is different between the

cell types. This significant difference can be due to very diverse patterns, with different biological interpretations. It is therefore necessary to classify interaction eQTLs based on these patterns. To achieve this classification, for every interaction eQTL we evaluated the strength of the effect in each cell type by calculating the difference between the mean expression of both genotypes. The cell type for which the effect was the strongest was labeled “High”. The cell type whose effect was most different from the strongest effect was labeled “Low”. The remaining two cell types were assigned to the group they resembled most closely. This classification allowed us to define 14 categories of interaction eQTLs. Additionally, we identified eQTLs that have a consistent effect across all four cell types. This category of consistent eQTLs consists of all probes satisfying the following three conditions: the gene has a significant main effect Q_i at marker m ; for the same marker m , the interaction $(Q \times CT)_{ij}$ is not significant; the mean eQTL effect across cell types has a coefficient of variation smaller than 0.3.

Estimating the FDR for the main QTL effect

We permuted the strain labels in the genotype data 100 times, maintaining the correlation of expression traits while destroying any genetic association. Then we applied the full ANOVA model and stored the genome-wide minimum p -value for each transcript. Based on the resulting empirical distribution of p -values, we estimated that a threshold of $-\log_{10} p = 6$ corresponds to a false discovery rate¹⁴ of 0.02 for the main QTL effect. The 99.9th percentile of the number of significant eQTLs per marker (i.e., the minimum size of statistically significant “eQTL hotspots”) is 28.

Estimating the FDR for interaction QTL effect

We estimated the residuals of the full ANOVA model after fitting all factors up to the main QTL effect at each marker for each transcript.¹⁵ Then we permuted the strain labels and applied the ANOVA model $y_{ij} = Q_i + CT_j + (Q \times CT)_{ij} + e_{ij}$ to the permuted residuals at each marker for each transcript and stored the genome-wide minimum p -value. Based on 100 permutations and the resulting empirical distribution of p -values, we estimated that a threshold of $-\log_{10} p = 6$ corresponds to a false discovery rate of 0.021 for interacting QTL effect. The 99.9th percentile of the number of significant eQTLs per marker (i.e., the minimum size of statistically significant “interaction hotspots”) is 8.

Detection of swapping eQTLs

Swapping eQTLs are those transcripts that show one eQTL in one cell type, but another eQTL in another cell type. From the full model mapping described above, we obtained 1283 transcripts with a significant interaction effect between genotype (first marker) and cell type. After taking into account the genetic

and interaction effects of the first marker, we scanned the genome excluding the region of the first marker (window size = 30cM) and tested if there was a significant interaction effect between genotype and cell type and whether this new interaction effect was classified in a different cell type category (see above Classification of eQTLs), which would indicate a swapping eQTL.

This means, for each transcript, a two-marker full model mapping was applied using the following model:

$$y_{ij} = \mu + CT_j + Q_i^* + (Q^* \times CT)_{ij} + Q_i + (Q \times CT)_{ij} + Q_i^* Q_i + e_{ij}$$

where y_{ij} is the gene's log intensity at the i th microarray ($i = 1, \dots, n$) and j th cell type; CT_j is the j th cell type effect; Q_i^* and $(Q^* \times CT)_{ij}$ are the main genotype effect at first marker and interaction effect between cell type and the genotype effect at this marker, where the first marker is defined as the marker with maximal interaction effect from previous one-marker full model mapping; Q_i is the genotype effect of the second marker; $(Q \times CT)_{ij}$ is the interaction effect between the i th genotype and j th cell type, $Q_i^* Q_i$ is the epistasis effect and e_{ij} is the residual error.

URLs

All raw data were deposited in the NCBI Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>, accession number GSE18067). All processed data were deposited in the GeneNetwork (www.genenetwork.org).

RESULTS

Genetic regulation of gene expression

We evaluated genome-wide RNA transcript expression levels in purified Lin⁻Sca-1⁺c-Kit⁺ multi-lineage cells, committed Lin⁻Sca-1⁻c-Kit⁺ progenitor cells, erythroid TER-119⁺ cells, and myeloid Gr-1⁺ cells, isolated from the bone marrow of ~25 genetically related and fully genotyped BXD – C57BL/6 (B6) X DBA/2 (D2) – recombinant inbred mouse strains.¹⁶ In this study, we exploit the fact that the purified cell populations are closely related, sometimes just a few cell divisions apart on the hematopoietic trajectory. The Lin⁻Sca-1⁺c-Kit⁺ cell population contains all stem cells with long-term repopulating ability, but also includes multipotent progenitors that still have lymphoid potential. Although long-term repopulating stem cells are known to only make up a fraction of the Lin⁻Sca-1⁺c-Kit⁺ population, for simplicity we will refer to this population as stem cells. The Lin⁻Sca-1⁻c-Kit⁺ cell population does not contain stem cells and lymphoid precursors, but does include common progenitors of the myeloid and erythroid lineages.¹⁷ Finally, TER-119⁺ cells and Gr-1⁺ cells are fully committed to the erythroid and myeloid lineages, respectively. Unsupervised clustering of the most varying transcripts

demonstrated that each of the four cell populations could easily be recognized based on expression patterns across all four cell types (Figure 1 and Table S1).

We observed strong and biologically significant variation in gene expression during hematopoietic differentiation, independent of mouse strain. However, the genetical genomics strategy, in which we focus on *inter*-strain gene expression differences, allows for a far more comprehensive understanding of the genetic regulatory links underlying this variation. QTL mapping of gene expression traits allows us to identify eQTLs; genomic regions that have a regulatory effect on those expression traits. Two types of eQTLs can be distinguished, i.e., those that map near (less than 10 Mb from) the gene which encodes the transcript (*local*) and those that map elsewhere in the genome (*distant*).¹⁸ Together, *local* and *distant* eQTLs constitute a genome-wide overview of the gene regulatory networks that are active in the cell type under study. The strongest eQTLs were found for genes that were expressed only in mouse strains carrying one specific parental allele, suggesting that local regulatory elements are distinct between the two alleles. Cases of such allele-specific expression included *H2-Ob* and *Apobec3*. These transcripts were only detectable in strains that carried the B6 allele of the gene (see Figures S1A–B). A global view of heritable variation in gene expression indicated that the strongest eQTLs are not associated with the most highly expressed genes, and that for most probes the expression difference between the B6 and D2 alleles is small (see Figures S1C–D).

Since the focus of this project is to study the influence of cellular differentiation state on regulatory links, we used ANOVA to distinguish between “*static*” eQTLs that show consistent genetic effects across the four cell types and “*dynamic*” eQTLs that are sensitive to cellular state (i.e., eQTLs that have a statistically significant genotype-by-cell-type interaction). We further partitioned *dynamic* eQTLs into different categories on the basis of their dynamics along the differentiation trajectory.

Cell-type-independent *static* eQTLs

The first eQTL category comprises genes that have *static* eQTLs across all four cell types under study. Variation in *Lxn* expression is shown as a representative example (Figure 2A, left panel). *Lxn* expression has previously been shown to be higher in B6 stem cells compared to D2 stem cells, and to be negatively correlated with stem cell numbers.¹⁹ In our dataset *Lxn* showed clear expression dynamics (it was most highly expressed in stem cells), and was indeed more strongly expressed in cells carrying the B6 allele, but the expression difference between mice carrying the B6 or D2 allele remained constant across all cell types.

In total, we identified 365 probes that displayed a *static* eQTL at threshold $p < 10^{-6}$ (FDR = 0.02). Among the 268 *locally*-regulated probes in this category was *H2-D1*. The histocompatibility gene *H2-D1* is known to be polymorphic between B6 and D2 mice, and would therefore be expected to be in the *static* eQTL category.

The remaining 97 probes mapped to *distant* eQTLs, i.e., their heritable expression variation was affected by the same *distant* locus in all four cell types (Table 1).

All probes that belonged to the *static* eQTL category are graphically depicted in an eQTL dot plot displaying the genomic positions of the eQTLs compared to the genomic positions of the genes by which the variably expressed transcripts were encoded (Figure 2A, right panel). Whereas in this plot *local* eQTLs appear on the diagonal, *distant* eQTLs appear elsewhere. In general, as has been reported before in eQTL studies, transcripts that were *locally* regulated showed strong linkage statistics. Not surprisingly, the statistical association between genotype and variation in transcript abundance for those transcripts that were controlled by *distant* loci was weaker. These genes are likely to be controlled by multiple loci, each contributing only partially to the phenotype, thereby limiting their detection and validation in the current experimental sample size. A list of all transcripts with significant *static* eQTLs is provided in Table S2.

Cell-type-dependent *dynamic* eQTLs

The second eQTL category comprises genes that have *dynamic* eQTLs across all four cell types under study. In total, we identified 1283 eQTLs ($p < 10^{-6}$, FDR = 0.021) that showed different genetic effects in different cell types, indicating that eQTLs are highly sensitive to cellular differentiation state (Table 1). Within this *dynamic* eQTL category, the first four subcategories are composed of eQTLs that were preferentially active in only one of the four cell types we analyzed (Figures 2B–E).

For example, *Slit2* mapped to a strong eQTL that was active only in stem cells. *Slit2* mRNA was only detected in the most primitive hematopoietic cell compartment in those BXD strains that carried the D2 allele at rs13478235, a SNP that mapped 629 kb away from the *Slit2* gene (Figure 2B, left panel). *Slit2* encodes an excreted chemorepellent molecule that is known to be expressed in embryonic stem cells,²⁰ to be involved in neurogenesis²¹ and angiogenesis,²² and to inhibit leukocyte chemotaxis.²³ We found a total of 140 genes that have eQTLs that are preferentially/selectively active in stem cells (Figure 2B, right panel, largest symbols, Table 1). These 140 genes included well-known candidate stem cell genes such as *Angpt1*, *Ephb2*, *Ephb4*, *Foxa3*, *Fzd6*, and *Hoxb5*. Interestingly, many transcripts with as yet unknown (stem cell) function were transcriptionally affected by stem-cell-specific eQTLs. Candidate novel stem cell genes include *Msh5*, and *Trim47*, in addition to a large collection of completely unannotated transcripts.

A total of 45, 531, and 295 eQTLs were found to be preferentially/selectively active in progenitors, erythroid cells, and myeloid cells, respectively (Table 1). Very distinct patterns of cell-type-specific gene regulation emerged when these eQTLs were visualized in genome-wide dot plots (Figures 2C–E). Using genome-wide p -value thresholds of $p < 10^{-6}$, we identified 53 *distantly*-regulated transcripts

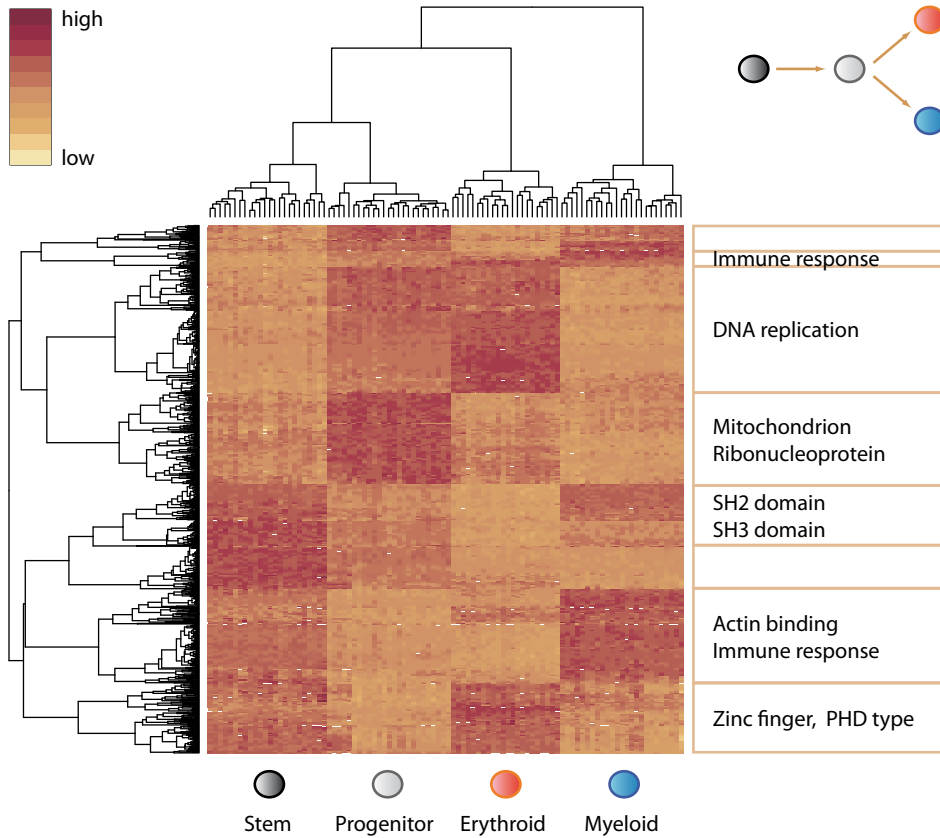


Figure 1. Mean expression levels for all probes in the four cell types. Unsupervised clustering including all probes for the 96 RNA samples follows cell-type (top hierarchical tree), while clustering of the 876 most varying probes reveals distinct categories of genes that show cell-type-specific expression (left hierarchical tree). The heat map shows the expression patterns of those probes and selected enriched gene categories in each major cluster. Discriminatory genes are enriched in various functional classes, including SH2/SH3 domain containing transcription factors for stem cells, mitochondrial genes for progenitor cells, genes involved in DNA replication and zinc fingers for erythroid cells, and immunoglobulin type genes for myeloid cells (all p -values < 0.05). For genes that belong to each of these clusters, see Table S1.

in stem cells, 13 in progenitor cells, 400 in erythroid cells, and 132 in myeloid cells. In erythroid and myeloid cells most of these transcripts mapped to relatively few genomic loci; these trans-bands are statistically significant, as assessed by a permutation approach taking expression correlation into account (see Methods).²⁴ Typically, transcripts mapping to a common marker showed a directional bias towards either B6 or D2 expression patterns.

In addition to the relatively simple eQTL dynamics that we have thus far illustrated, more complex eQTL dynamics were also detected using this

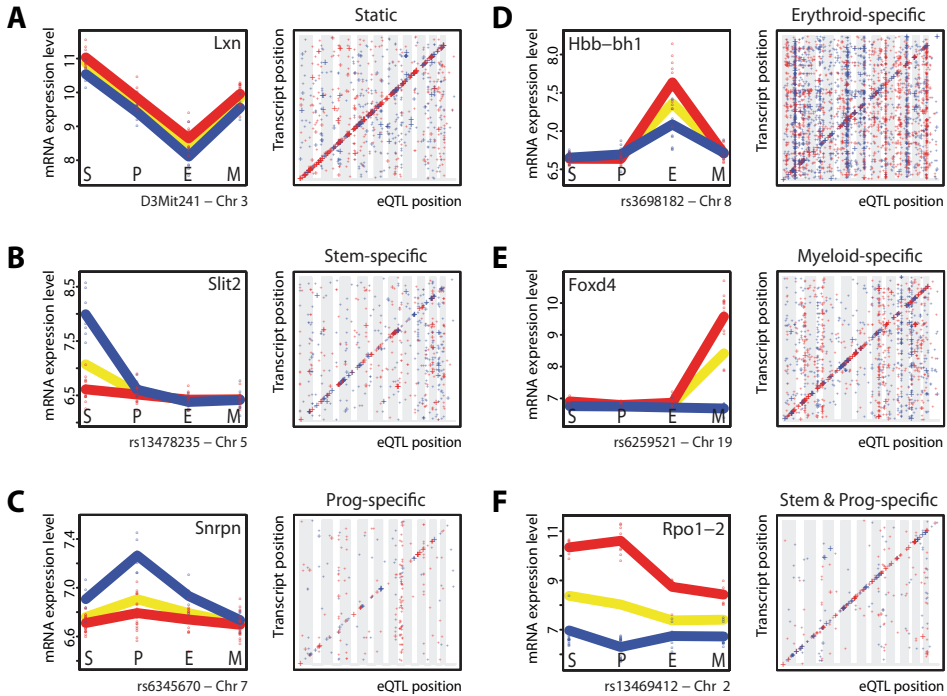


Figure 2. Identification of static and dynamic eQTLs. (A) Genome-wide identification of cell-type-independent *static* eQTLs. (Left panel) *Lxn* mRNA levels were analyzed in all 4 cell types. Each circle represents an individual sample (strain). The yellow line shows mean expression levels across all strains. The red and blue lines indicate mean *Lxn* expression levels in strains that carry the B6 or D2 *Lxn* allele, respectively. The genetic effect of parental alleles on *Lxn* expression levels was consistent in all cell types. (Right panel) Individual probes that detected a transcript that was consistently controlled by the same eQTL in all 4 cell types. The y-axis indicates the physical position of the encoding gene, the x-axis provides the genomic position of the marker with strongest linkage statistics. Vertical gray and white bandings indicate different chromosomes, ranging from chromosome 1 to X. The size of each symbol reflects the strength of the genetic association: eQTLs with p -values $< 10^{-8}$ are represented by the largest crosses, p -values between 10^{-6} and 10^{-8} are shown with medium crosses, while small crosses refer to eQTLs with p -values between 10^{-4} and 10^{-6} . The color coding (red and blue) indicates the parental allele of the eQTL that caused a higher gene expression (B6 is red and D2 is blue). (B–E) Genome-wide identification of transcripts that are controlled by cell-type-specific eQTLs. (Left panels) Expression data for some transcripts that were affected by cell-type-specific eQTLs (B: *Slit2* in stem cells, C: *Snrpn* in progenitor cells, D: *Hbb-bh1* in erythroid cells and E: *Foxd4* in myeloid cells). (Right panels) Genome-wide distribution of eQTLs that were preferentially/uniquely detected in each of the four cell populations. (F) Transcripts that were controlled by eQTLs in both stem and progenitor cells. An example is *Rpo1-2*. Full lists of all genes belonging to the eQTL (sub)categories shown here are provided in Table S2.

Table 1. Overview of static and dynamic eQTLs ($p < 10^{-6}$): number of probes and associated markers.

eQTL category	eQTL subcategory		# probes	# markers	# probes / # markers
<i>Static</i>	All	<i>Local</i>	268	161	1.66
		<i>Distant</i>	97	76	1.28
		Total	365	213	1.71
<i>Dynamic</i>	All	<i>Local</i>	642	282	2.28
		<i>Distant</i>	641	276	2.32
		Total	1283	445	2.88
	Stem-specific	<i>Local</i>	87	66	1.32
		<i>Distant</i>	53	42	1.26
		Total	140	105	1.33
	Progenitor-specific	<i>Local</i>	32	27	1.19
		<i>Distant</i>	13	12	1.08
		Total	45	39	1.15
	Erythroid-specific	<i>Local</i>	131	90	1.46
		<i>Distant</i>	400	164	2.44
		Total	531	223	2.38
	Myeloid-specific	<i>Local</i>	163	121	1.35
		<i>Distant</i>	132	72	1.83
		Total	295	179	1.65

approach. For example, *Rpo1-2* is a transcript that shows a strong *local* eQTL in the two non-committed lineages included in our study, but shows a much weaker genetic effect in erythroid and myeloid cells (Figure 2F). Whereas in mice carrying the B6 allele of *Rpo1-2* the overall expression of the gene decreased substantially during differentiation of progenitor to erythroid cells, in mice carrying the D2 allele expression slightly increased. This observation hints at complex regulatory mechanisms underlying the expression of this gene. Full lists of genes in each *dynamic* eQTL subcategory described thus far are supplied in Table S2. Additional subcategories and their exact definitions are explained more extensively in the Methods section, and complete results of all *dynamic* eQTLs are available in Table S3.

Detailed analysis of *static* and *dynamic* eQTLs

eQTL dynamics can be caused by transcription factors being switched on/off upon cellular differentiation, or by a transcription factor showing changed specificity due to variations in regulatory input. We found that most (>75%) of the *dynamic* eQTLs are active in only one of the four cell types under study (Figure 3A).

A more detailed analysis 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 (Figure 3B). Next, we explored whether we could find transcripts that were regulated by distinct eQTLs in different cell types (see Methods). Such eQTL “swapping” would indicate major changes in transcriptional regulatory networks. We could find no evidence for such cases. However, given our limited population size we have a low power to detect multiple eQTLs, so swapping eQTLs may still exist but remain undetected in our experimental setting.

It has been described that not all *local* eQTLs in genetical genomics experiments reflect actual expression differences between mouse strains, but rather indicate differential hybridization caused by polymorphisms in the sequences recognized by the probes.²⁵ For this reason, we divided both the *static* and *dynamic* eQTL categories in *local* and *distant* eQTLs, and indicated the number of probes that hybridized to sequences that are known to contain polymorphisms (Figure 3C). As expected, the *static* eQTL category contained a higher number of such potential false *local* eQTLs. If these false positive eQTLs could be removed, the relative abundance of *dynamic* eQTLs would be higher, indicating that our study may even conservatively underestimate the level of eQTL dynamics.

DISCUSSION

We found that many eQTLs are highly sensitive to the developmental state of the cell population under study. Even when the purified cells were only separated by a few cell divisions, eQTLs demonstrated a remarkable plasticity. Furthermore, we provide evidence that the cell-stage-sensitivity of eQTLs is often intertwined with gene expression variation during development. We did not identify target genes that were regulated by distinct eQTLs in different cell types, suggesting that large-scale changes within transcriptional regulatory networks are not common.

The fact that eQTLs appear to be highly cell-type-dependent highlights the importance of using well-characterized purified cell types in eQTL studies. In particular, eQTL studies of physiological or disease processes²⁶⁻²⁹ should target the relevant cell type as precisely as possible, i.e. they should use cells or tissues directly involved in the patho-physiological process. This could even mean that several different cell types need to be separately studied, in particular if developmental trajectories are affected.³⁰ Using unfractionated bone marrow cells, we would have missed many of the diverse and dynamic patterns that we uncovered here, both at the expression level and at the genetic regulatory level. Even so, the four cell populations that we studied are still heterogeneous and further subfractionation of these populations based on different sets of markers would have resulted in even more precise regulatory maps.

Many genetical genomics experiments have used highly heterogeneous samples, in which mRNA from a variety of different cell types was pooled.^{4;5;31-34}

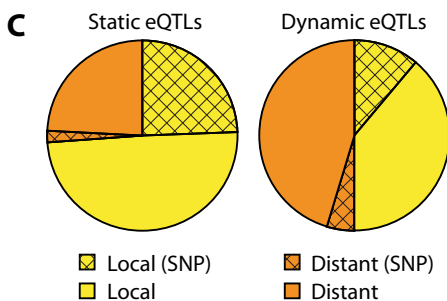
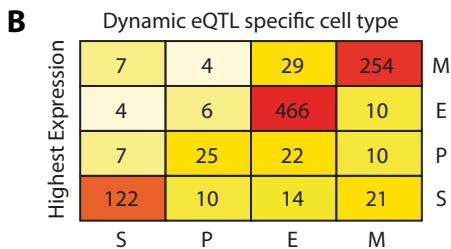
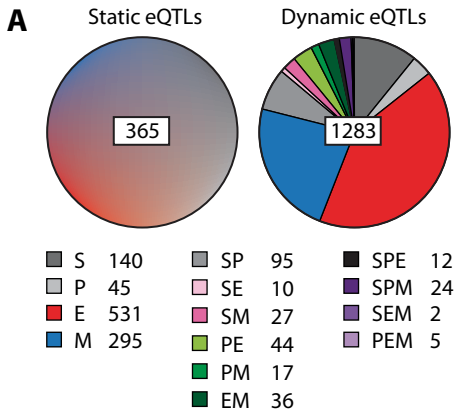


Figure 3. Quantitative overview of static and dynamic eQTLs. (A) Pie charts presenting all 365 static and 1283 dynamic eQTLs that were detected with $p < 10^{-6}$. Dynamic eQTLs are subdivided in all 14 categories of interaction eQTLs. (B) Matrix showing the four cell-type-dependent dynamic eQTL categories and the cell type in which the gene was expressed most highly. (C) All static and dynamic eQTLs are subdivided in local and distant eQTLs. Shown is which number of eQTLs was detected by Illumina probes that hybridize to sequences that are known to contain polymorphisms (SNPs) between the two parental strains. Abbreviations: S, stem cells; P, progenitor cells; E, erythroid cells; M, myeloid cells.

In such mixed samples it is usually impossible to ensure that the contribution of individual cell types to the mixture is the same across samples. As a result, important parts of the variation in gene expression could arise from different sample compositions. For example, if in whole brain samples a heritable morphological or developmental trait leads to an increased size of some brain regions, this can cause apparent hotspots for transcripts that are specific for those particular regions. Our data provide a valuable tool for studying the exact consequences of sample heterogeneity on eQTL mapping: a further study could simulate a collection of samples made of computed mixtures of different hematopoietic cells in defined proportions. Clearly, cell purification strategies are essential to identify those cell-type-specific eQTLs that would otherwise be “masked” in heterogeneous cell

populations. Therefore, future genetical genomics studies should be realized on as many cell types or cellular differentiation states as possible, and ideally even on the scale of individual cells.

All data presented in this paper were deposited in the online database GeneNetwork (www.genenetwork.org), an open web resource that contains genotypic, gene expression, and phenotypic data from several genetic reference populations of multiple species (e.g. mouse, rat and human) and various cell types and tissues.^{35,36} It provides a valuable tool to integrate gene networks and phenotypic traits, and also allows cross-cell type and cross-species comparative gene expression and eQTL analyses. Our data can aid in the identification of candidate modulators of gene expression and/or phenotypic traits,³⁷ and as such can serve as a starting point for hypothesis-driven research in the fields of stem cell biology and hematology.

ACKNOWLEDGMENTS

We thank Guus Smit and Sabine Spijker for providing BXD mice; Geert Mesander and Henk Moes for assistance in cell sorting; and Arthur Centeno and Rob W. Williams for depositing our data in www.genenetwork.org. This work was supported by a Horizon grant from the Netherlands Genomics Initiative (050-71-055); a Biorange grant SP1.2.3 from the Netherlands Genomics Initiative/ Netherlands Bioinformatics Centre; two VICI grants from the Netherlands Organization for Scientific Research (NWO) to G.d.H (918-76-601) and R.C.J (865-04-001); and by grants from the European Community (Marie Curie RTN EUrythron, MRTN-CT-2004-005499; and EuroSystem, 200720). X.W. is supported by the National Institutes of Health (U01-AA014425 and P20-DA21131).

SUPPORTING INFORMATION AVAILABLE ONLINE

Figure S1. Analysis of the quantitative aspects of eQTLs

Table S1. Clustering results

Table S2. Principal eQTL categories

Table S3. All *dynamic* eQTLs

REFERENCES

- Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Genet.* 2001;17(7):388-391.
- Schadt EE, Monks SA, Drake TA et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 2003;422(6929):297-302.
- Bystrykh L, Weersing E, Dontje B et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat.Genet.* 2005;37(3):225-232.
- Chesler EJ, Lu L, Shou S et al. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* 2005;37(3):233-242.
- Hubner N, Wallace CA, Zimdahl H et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat.Genet.* 2005;37(3):243-253.
- Petretto E, Mangion J, Dickens NJ et al. Heritability and tissue specificity of expression quantitative trait loci. *PLoS. Genet.* 2006;2(10):e172.
- Monks SA, Leonardson A, Zhu H et al. Genetic inheritance of gene expression in human cell lines. *Am.J.Hum.Genet.* 2004;75(6):1094-1105.
- Morley M, Molony CM, Weber TM et al. Genetic analysis of genome-wide variation in human gene expression. *Nature* 2004;430(7001):743-747.
- Ivanova NB, Dimos JT, Schaniel C et al. A stem cell molecular signature. *Science* 2002;298(5593):601-604.
- Chambers SM, Boles NC, Lin KY et al. Hematopoietic Fingerprints: An Expression Database of Stem Cells and Their Progeny. *Cell Stem Cell* 2007;1(5):578-591.
- Kiel MJ, Yilmaz OH, Iwashita T et al. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* 2005;121(7):1109-1121.
- Forsberg EC, Prohaska SS, Katzman S et al. Differential expression of novel potential regulators in hematopoietic stem cells. *PLoS.Genet.* 2005;1(3):e28.
- Dennis G, Jr., Sherman BT, Hosack DA et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 2003;4(5):3.
- Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc.Natl. Acad.Sci.U.S.A* 2003;100(16):9440-9445.
- Anderson MJ, Ter Braak CJF. Permutation tests for multi-factorial analysis of variance. *J Stat Comput Simul.* 2003;73:85-113.
- Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC.Genet.* 2004;57.
- Bryder D, Rossi DJ, Weissman IL. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *Am.J.Pathol.* 2006;169(2):338-346.
- Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat.Rev.Genet.* 2006;7(11):862-872.
- Liang Y, Jansen M, Aronow B, Geiger H, Van Zant G. The quantitative trait gene latexin influences the size of the hematopoietic stem cell population in mice. *Nat.Genet.* 2007;39(2):178-188.
- Katoh Y, Katoh M. Comparative genomics on SLIT1, SLIT2, and SLIT3 orthologs. *Oncol.Rep.* 2005;14(5):1351-1355.
- Wang KH, Brose K, Arnott D et al. Biochemical purification of a mammalian slit protein as a positive regulator of sensory axon elongation and branching. *Cell* 1999;96(6):771-784.
- Wang B, Xiao Y, Ding BB et al. Induction of tumor angiogenesis by Slit-Robo signaling and inhibition of cancer growth by blocking Robo activity. *Cancer Cell* 2003;4(1):19-29.
- Wu JY, Feng L, Park HT et al. The neuronal repellent Slit inhibits leukocyte chemotaxis induced by chemotactic factors. *Nature* 2001;410(6831):948-952.
- Breitling R, Li Y, Tesson BM et al. Genetical genomics: spotlight on QTL hotspots. *PLoS.Genet.* 2008;4(10):e1000232.
- Alberts R, Terpstra P, Li Y et al. Sequence polymorphisms cause many false cis eQTLs. *PLoS.ONE.* 2007;2(7):e622.
- Schadt EE, Lamb J, Yang X et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat.Genet.* 2005;37(7):710-717.
- Goring HH, Curran JE, Johnson MP et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat.Genet.* 2007;39(10):1208-1216.

28. Emilsson V, Thorleifsson G, Zhang B et al. Genetics of gene expression and its effect on disease. *Nature* 2008;452(7186):423-428.
29. Chen Y, Zhu J, Lum PY et al. Variations in DNA elucidate molecular networks that cause disease. *Nature* 2008;452(7186):429-435.
30. Li Y, Breitling R, Jansen RC. Generalizing genetical genomics: getting added value from environmental perturbation. *Trends Genet.* 2008;24(10):518-524.
31. Li Y, Alvarez OA, Gutteling EW et al. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS.Genet.* 2006;2(12):e222.
32. West MA, Van Leeuwen H, Kozik A et al. High-density haplotyping with microarray-based expression and single feature polymorphism markers in *Arabidopsis*. *Genome Res.* 2006;16(6):787-795.
33. Keurentjes JJ, Fu J, Terpstra IR et al. Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc.Natl.Acad. Sci.U.S.A* 2007;104(5):1708-1713.
34. Whiteley AR, Derome N, Rogers SM et al. The phenomics and expression quantitative trait locus mapping of brain transcriptomes regulating adaptive divergence in lake whitefish species pairs (*Coregonus* sp.). *Genetics* 2008;180(1):147-164.
35. Chesler EJ, Lu L, Wang J, Williams RW, Manly KF. WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nat. Neurosci.* 2004;7(5):485-486.
36. Wang J, Williams RW, Manly KF. WebQTL: web-based complex trait analysis. *Neuroinformatics.* 2003;1(4):299-308.
37. Gerrits A, Dykstra B, Otten M, Bystrykh L, De Haan G. Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny. *Immunogenetics* 2008;60(8):411-422.

CHAPTER

4

INFERRING COMBINATORIAL ASSOCIATION LOGIC NETWORKS IN MULTIMODAL GENOME-WIDE SCREENS

Jeroen de Ridder, Alice Gerrits, Jan Bot,
Gerald de Haan, Marcel Reinders and Lodewyk Wessels

Bioinformatics. 2010 Jun 15;26(12):i149-57

ABSTRACT

Motivation: We propose an efficient method to infer combinatorial association logic networks from multiple genome-wide measurements from the same sample. We demonstrate our method on a genetical genomics dataset, in which we search for Boolean combinations of multiple genetic loci that associate with transcript levels.

Results: Our method provably finds the global solution and is very efficient with runtimes of up to four orders of magnitude faster than exhaustive search. This enables permutation procedures for determining accurate false positive rates and allows selection of the most parsimonious model. When applied to transcript levels measured in myeloid cells from 24 genotyped recombinant inbred mouse strains, we discovered that nine gene clusters are putatively modulated by a logical combination of trait loci rather than a single locus. A literature survey supports and further elucidates one of these findings. Due to our approach, optimal solutions for multi-locus logic models and accurate estimates of the associated False Discovery Rates become feasible. Our algorithm therefore offers a valuable alternative to approaches employing complex, albeit sub-optimal optimization strategies to identify complex models.

INTRODUCTION

To explain complex biological phenomena it is of vital importance to measure – in the same sample – all relevant (complementary) biological variables, and to measure these at a genome-wide scale. For this reason, many *multimodal* screens have been performed that have complemented transcriptional profiling with, among others, copy number variation measurements, transcription factor binding assays, methylation status profiling or genotype calls.¹⁻⁴

A common aim in analyzing these multimodal datasets is to find associations between the biological variables measured to infer their regulatory role. Consider, for instance, a study in which expression profiles and genome-wide genotype data were obtained in hematopoietic cells from a panel of fully homozygous recombinant inbred mouse strains (Figure 1A). This ‘genetical genomics’ approach enables the determination of expression quantitative trait loci (eQTLs) characterized by strong associations between the genotype and the observed expression levels.^{5,6} In the absence of a strong direct association between the genotype and gene expression, real multi-locus interactions may still be present, due to epistatic interaction.^{7,8} Such interactions may not be detectable as (marginal) direct associations between the genotype and gene expression (Figure 1B).

To alleviate this, approaches which evaluate the joint association of multiple loci and a phenotype of interest are required. Several approaches have been proposed to attack this problem. These approaches differ mostly regarding the way the associations are modeled and the strategy employed to solve the combinatorial optimization problem. Some approaches follow what could be loosely termed a two-stage approach, where all two-locus models are first evaluated, which, in stage two, are used in a greedy search to yield multi-locus models.^{9,10} Approaches employing more advanced strategies to traverse the space of possible models are represented by a genetic programming approach¹¹ and Markov Chain Monte Carlo (MCMC) approaches associated with Bayesian analyses.^{12,13} Since two-stage approaches have been demonstrated to be suboptimal¹⁴ and advanced search strategies such as MCMC are very sensitive to their implementation and parameter settings and are not guaranteed to be optimal, an approach which finds a provably global solution to a selected model within reasonable time is highly desirable. Of particular interest is the method of Ljungberg *et al*¹⁵, which is used for the Pair-Scan analysis that is available on The GeneNetworkⁱ. Ljungberg *et al.* stress the importance of performing a global search rather than relying on greedy searches by (pre)selecting markers based on their marginal effects. To deal with the computational complexity associated with such an optimization problem, the authors present a method to find global optima of a linear regression problem for up to three predictors that is fast enough to be employed in permutation procedures.

ⁱ <http://genenetwork.org>

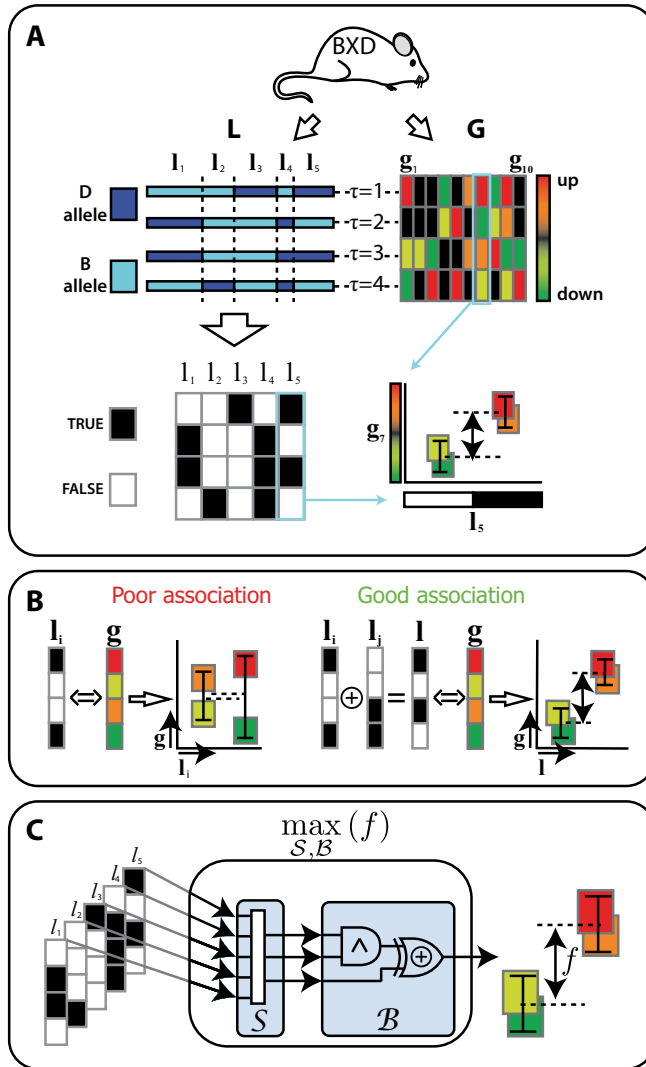


Figure 1. Schematic overview of data and association inference. (A) A panel of BXD mice that is densely genotyped and expression profiled. The genotype data can be considered as binary vectors by choosing a binary encoding of the alleles (in the figure D=TRUE and B=FALSE) and putting thresholds that divide the genome into loci such that each locus differs in at least one element from its neighbors. The cartoon shows that good association is obtained between Locus 5 and Gene 7 because elevated expression is consistently observed in conjunction with the D allele of Locus 5. (B) Interaction among genetic features may destroy direct associations between individual loci and genes. The cartoon shows that configurations exist in which the gene expression can only be predicted by considering two loci simultaneously (using Boolean XOR logic). (C) By inferring CAL networks, interaction among genetic features is taken into account in the association inference. Inferring CAL networks is achieved by selecting the input loci with the selection function S and combining these with the appropriate Boolean function B , such that the association (as measured by a scoring function f) between the network output and the gene of interest is maximized.

In contrast to the class of additive models employed by Ljungberg *et al.* (and many other approaches), we follow others^{11,12,16} and employ Boolean combinatorial logic to explicitly incorporate interactions in the eQTL inference. To this end, we infer combinatorial association logic (CAL) networks that combine the observed genotypes through AND (\wedge), OR (\vee) and XOR (\oplus) functions by searching for associations between the result of the Boolean operation and the gene expression. The Boolean AND function can be used if altered expression is consistently observed in combination with a particular combination of two alleles (which do not necessarily have to be equal), but remains unchanged in all other genotype configurations. An example of a situation in which this may be observed is the case of two parallel pathways that only promote transcription of their downstream target when the genes in these pathways have specific alleles. Conversely, we may also consistently observe differential transcription in the strains for which either one of two loci is of a certain genotype. This may, for instance, be observed in case of a cascaded signaling pathway: a silencing mutation in one of the alleles can repress the entire pathway, regardless of which gene in the cascade contained this mutation. Boolean OR (\vee) and XOR (\oplus) are capable of capturing this behavior (Figure 1B).

Like the search for optimal predictors in the additive model, inferring optimal predictors of a Boolean function is a challenging computational problem, especially considering that more complex combinations of these functions are also possible. Moreover, we noted that the objective function that needs to be optimized is highly discontinuous and nonlinear so that standard optimization techniques, such as Genetic algorithms, Simulated Annealing and MCMC do not provide an optimal solution. Nevertheless, an efficient and –most importantly – global solution is highly desirable, since this allows permutation procedures with which significance estimates of the discovered associations can be realized.¹⁵

In the following we will mathematically prove that, under reasonable conditions, CAL network inference provides an efficient way to obtain globally optimal multi-locus models that associate multiple genomic loci with the expression of target genes. We illustrate our approach on the genetical genomics dataset from Gerrits *et al.*,¹⁷ and using these data show that 100% accuracy is achieved at runtimes that are a fraction of those required for exhaustive search. Furthermore, we observe that using this approach complex associations are revealed that otherwise would have gone unnoticed. As such, our approach offers a useful alternative to the commonly used additive models and suboptimal search strategies.

METHODS

CAL network search

The construction of a CAL network that predicts the expression profile from a set of binary predictors can be formulated as an optimization problem. Interesting logic

networks are those for which maximal association between the network output and the gene expression is obtained. Let \mathbf{g} be the $(T \times 1)$ vector, with T the number of samples, containing the expression values of a gene, and \mathbf{L} the $(T \times L)$ matrix of binary predictors, e.g. the genotypes, where L is the number of predictors. A CAL network \mathcal{L} is defined in terms of $\mathcal{S}(\mathbf{L}; \mathbf{n}): \mathbb{B}^L \rightarrow \mathbb{B}^N$, a selection function that selects N columns from \mathbf{L} , and $\mathcal{B}(\mathbf{I}): \mathbb{B}^N \rightarrow \mathbb{B}$, a Boolean logic function that specifies the network topology. In the latter, $(T \times N)$ matrix \mathbf{I} is a concatenation of the columns selected by \mathcal{S} , i.e. $\mathbf{I} = (\mathbf{i}_{n(1)}, \dots, \mathbf{i}_{n(N)})$, where \mathbf{n} is a $(N \times 1)$ vector containing the indices of the selected columns. Consequently, CAL network \mathcal{L} maps the genotype matrix \mathbf{L} to a $(T \times 1)$ output vector \mathbf{y} as follows:

$$\mathbf{y} = \mathcal{L}(\mathbf{L}; \mathcal{B}, \mathbf{n}) = \mathcal{B}(\mathcal{S}(\mathbf{L}; \mathbf{n})). \quad (1)$$

The association between \mathbf{g} and \mathbf{y} is quantified with an association measure $f(\mathbf{g}, \mathbf{y})$:

$$f(\mathbf{g}, \mathbf{y}) = \begin{cases} \frac{|\bar{x}_0 - \bar{x}_1|}{\sqrt{\frac{(n_0-1)s_0^2 + (n_1-1)s_1^2}{n_0+n_1-2} \left(\frac{1}{n_0} + \frac{1}{n_1}\right)}} & \text{if } (n_0 > \eta) \cup (n_1 > \eta) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

For notational convenience, we used $\mathbf{x}_0 = \{\mathbf{g}(\tau): \mathbf{y}(\tau) = 0, \forall \tau \in (1, \dots, T)\}$ and $\mathbf{x}_1 = \{\mathbf{g}(\tau): \mathbf{y}(\tau) = 1, \forall \tau \in (1, \dots, T)\}$, i.e. vector \mathbf{g} is split into \mathbf{x}_0 and \mathbf{x}_1 according to the Boolean values in \mathbf{y} . Furthermore, \bar{x}_0 (\bar{x}_1), s_0^2 (s_1^2) and n_0 (n_1) are defined as the sample mean, the sample variance and the number of elements in \mathbf{x}_0 (\mathbf{x}_1), respectively. Note that Equation (2) is equal to the absolute value of the t -statistic, except when n_0 or n_1 becomes too small, which ensures high f -values are only obtained in case \mathbf{x}_0 and \mathbf{x}_1 have at least η elements.

The inference of CAL networks is a computationally challenging problem. Primarily, because the feature selection problem, i.e. finding the optimal vector \mathbf{n} , critically depends on the number of features that are considered. In the case of genetic markers, this easily runs in the several hundreds to thousands. Moreover, the optimal subset of markers is heavily dependent on how these markers are combined, i.e. dependent on the optimal Boolean function \mathcal{B} . Altogether, one frequently has to rely on greedy search strategies that easily get stuck in local optima or near exhaustive searches that are computationally too expensive, especially when employed in permutation procedures required to assess statistical significance.

Our solution to this problem hinges upon two observations. First, in most practical datasets the sample size is relatively small, especially when compared to the number of features. This means that we can limit ourselves to considering only small CAL networks with few inputs, since larger networks are prone to overfitting, which makes them less informative. For this reason, and because most networks have many equivalent topologies that do not need to be evaluated due to symmetry, the set containing all unique and meaningful network topologies

$\{\mathcal{B}_j; j=1,2,\dots\}$ is relatively small (in the order of 10-100, depending on the desired topology). Consequently, the set of optimal input vectors $\{\mathbf{n}_j^*; j=1,2,\dots\}$, associated with each \mathcal{B}_j , can be found by fixing \mathcal{B}_j and maximizing for each \mathcal{B}_j separately:

$$\mathbf{n}_j^* = \underset{\mathbf{n}}{\operatorname{argmax}} \{f(\mathbf{g}, \mathcal{B}_j; \mathcal{S}(\mathbf{L}; \mathbf{n}))\}. \quad (3)$$

Second, we observe that Equation (3) still represents a complex optimization problem that can be significantly simplified by employing an approximation to the association measure, denoted by \hat{f} . In the following, we show that maximizing \hat{f} is equivalent to maximizing f , but the maximization of the former can be very efficiently realized by using a branch and bound search. Before defining \hat{f} we define the Boolean vector \mathbf{y}^{opt} as the solution for which f reaches a global maximum independent of the network topology, i.e. $\mathbf{y}^{\text{opt}} = \underset{\mathbf{y}}{\operatorname{argmax}} f(\mathbf{g}, \mathbf{y})$. Note that \mathbf{y}^{opt} can be easily determined by sorting the gene expression vector \mathbf{g} and evaluating all positions for a threshold t that splits \mathbf{g} into x_0 and x_1 (Figure 2A). For \hat{f} , we use the weighted Hamming similarity between \mathbf{y}^{opt} and the network output \mathbf{y} :

$$\hat{f}(\mathbf{y}^{\text{opt}}, \mathbf{y}) = \sum_{\forall \tau} w(\tau) I(y^{\text{opt}}(\tau) = y(\tau)) \quad (4)$$

where $w(\tau) > 0 \forall \tau$ denotes the weight for sample τ , and $I(\cdot)$ is the indicator function, evaluating to '1' if the τ^{th} element of vectors \mathbf{y}^{opt} and \mathbf{y} are equal.

For an example gene expression vector, Figure 2B shows 500 random samples of (\hat{f}, f) -pairs, in case all weights are equal to one. Although the trend of this distribution is monotonically increasing, the spread around the trend is substantial. This is undesirable because a maximum in \hat{f} is only guaranteed to correspond to a maximum in f in case there is a direct one-to-one relation between them. Clearly, this is not the case in Figure 2B, since each value of \hat{f} corresponds to many values of f . However, by optimizing the weights such that the difference between \hat{f} and f is minimal, a near one-to-one relation can be obtained, as exemplified by Figure 2C. With the proper adjustments, detailed below, it is thus ensured that maximizing \hat{f} is equivalent to maximizing f . The major advantage of maximizing \hat{f} instead of f is that in the former each sample has an independent contribution to the association measure. This can be readily exploited using a branch and bound search, so that it is possible to avoid the expensive evaluation of the association measure.

Optimizing Equation (3)

Here, we show that optimizing Equation (3) can be achieved by first determining $\hat{f}^* = \max_{\mathbf{n}}(\hat{f})$, where \hat{f} was defined in Equation (4). After this the search for $f^* = \max_{\mathbf{n}} f$ is readily solved by searching in the neighborhood of \hat{f}^* .

For a single sample τ , let I^{τ} be the set of input combinations such that $y(\tau) = y^{\text{opt}}(\tau) \forall \mathbf{n} \in I^{\tau}$, where $\mathbf{y} = \mathcal{L}(\mathbf{L}; \mathcal{B}, \mathbf{n})$.ⁱⁱ Figures 3A-C show how I^{τ} can be inferred

ⁱⁱ Since we optimize Equation (3) for each \mathcal{B}_j separately, we omit its subscript if its meaning is inconsequential.

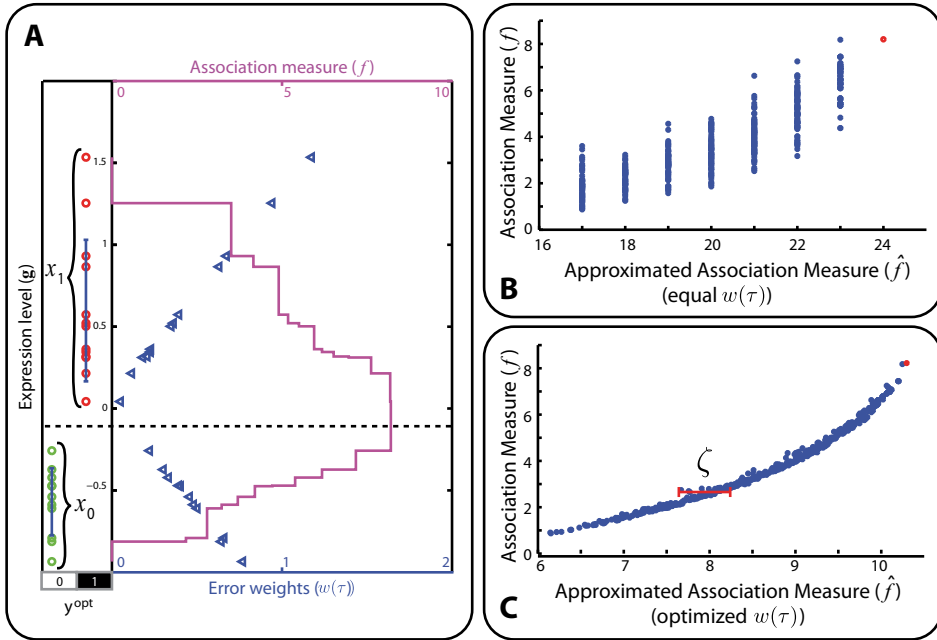


Figure 2. Association versus approximated association. (A) Example gene expression vector (circles) split in x_0 and x_1 according to y^{opt} . The magenta line denotes the association measure f , defined in Equation (2), as a function of a threshold t that splits the expression vector in x_0 and x_1 . The blue triangles indicate the error weights $w(\tau)$ that result after optimizing them. (B and C) 500 random samples that are generated by introducing up to seven bit-flips in y^{opt} to show the relation between \hat{f} and f . The red dot indicates \hat{f} and f values for y^{opt} . (B) shows the samples in case the weights are assumed equal. Although the trend of the data is monotonically increasing, a large spread around this trend is observed. (C) shows the same samples in case the weights are optimized, resulting in a near one-to-one relation between \hat{f} and f .

from \mathbf{L} and the truth table of \mathcal{B} . For a set of samples C , the input combinations $\mathbf{n} \in I^{(C)}$ for which all $\tau \in C$ reach the optimal output \mathbf{y}^{opt} are found by taking the intersection of all the individual sets of input combinations, i.e. $I^{(C)} = \bigcap_{\tau \in C} I^{(\tau)}$. Note that, under the assumption that each sample has at least one non-zero locus, $I^{(\tau)} \neq \emptyset \forall \tau$. In other words, for individual samples there always exists a combination of inputs for which the network can reach the desired optimal output \mathbf{y}^{opt} . However, for an arbitrary combination of samples this is clearly not the case. If we observe that $I^{(C)} = \emptyset$, this means that for the collection of samples in C there does not exist a valid combination of inputs. Moreover, if $I^{(C)} = \emptyset$, all supersets of C will also result in the empty set. Finally we note that, by choosing a convenient binary encoding, $I^{(\tau)}$ and $I^{(C)}$ can be computed very efficiently by means of bitwise XNOR and AND operations, respectively (see Figure 3D and Figure S1 for details).

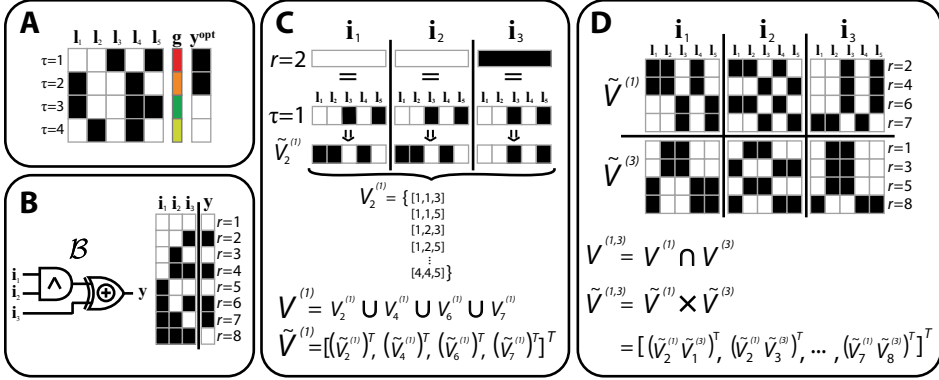


Figure 3. Computation of solution sets for each sample. (A) Example data from Figure 1A. (B) The topology and the truth table of the Boolean function \mathcal{B} under investigation. (C) Explanation by example of the calculation of $V^{(\tau)}$, the set of all possible input combinations to \mathcal{B} such that $y^{\text{opt}}(\tau)=y(\tau)$. This panel shows how $V^{(1)}$ is determined. Since $y^{\text{opt}}(1)=1$, the rows from the truth table for which $y=1$ are applicable, i.e., $r=\{2,4,6,7\}$. According to $r=2$, the desired output for $\tau=1$ is obtained by selecting any of the loci that are ‘0’ for inputs i_1 and i_2 , and loci that are ‘1’ for input i_3 . Accordingly, for i_1 we may select from the set: $\{i_1, i_2, i_4\}$. This can be efficiently calculated by taking the XNOR (evaluates to ‘1’ when both inputs are equal) between row $\tau=1$ from the data matrix and the row $r=2$ from the truth table, as shown in (C). Observe that the result is an efficient encoding of all the possible input combinations that satisfy $y^{\text{opt}}(1)$ while using $r=2$ from the truth table. In general, we denote this set by $V_r^{(\tau)}$, and its binary encoding by $\tilde{V}_r^{(\tau)}$. To determine the complete set of valid input combinations for $\tau=1$, rows 4, 6 and 7 need to be considered in a similar fashion. $V^{(1)}$ is now determined by taking the union of the subsets, i.e. $V^{(1)}=V_2^{(1)} \cup V_4^{(1)} \cup V_6^{(1)} \cup V_7^{(1)}$, which, in binary form, may be represented by a concatenation of $\tilde{V}_2^{(1)}, \tilde{V}_4^{(1)}, \tilde{V}_6^{(1)}$ and $\tilde{V}_7^{(1)}$. (D) This panel shows the valid input combinations for $\tau=1$ and $\tau=3$ in binary representation (i.e. $\tilde{V}^{(1)}$ and $\tilde{V}^{(3)}$). For any set of samples C the input combinations for which the output equals y^{opt} can be obtained by taking the intersection of the individual sets. In binary representation, this is equivalent to taking the row-wise Cartesian product (row-wise product of all combinations of rows), as is shown in the panel.

With these definitions in mind, we propose the following lemma:

LEMMA 1.

$$\hat{f}^* = \max_C \sum_{\forall \tau \in C} w(\tau) \quad \text{subject to: } V^{(C)} \neq \emptyset \quad (5)$$

PROOF. Let $C^* = \text{argmax}_C \sum_{\forall \tau \in C} w(\tau)$, i.e. C^* is the set of solutions for which \hat{f}^* is obtained. Since it is required that $V^{(C^*)} \neq \emptyset$, there must be at least one solution \mathbf{n} such that $y^{\text{opt}}(\tau) = y(\tau) \forall \tau \in C^*$. Since for C^* the optimum in \hat{f} is obtained, it must also hold that $y^{\text{opt}}(\tau) \neq y(\tau) \forall \tau \notin C^*$. This means that Equation (4) can be rewritten as follows: $\sum_{\forall \tau} w(\tau) I(y^{\text{opt}}(\tau) = y(\tau)) = \sum_{\forall \tau \in C^*} w(\tau)$, proving the statement in this lemma. ■

As argued by Lemma 1, Equation (4) is thus maximized by having as many samples in C as possible, while taking into account their respective weights $w(\tau)$.

Before we will show that Equation (5) fits a branch and bound framework, we first make the observation that for the relation between \hat{f} and f the following holds:

$$(\hat{f}(\mathbf{y}^{\text{opt}}, \mathbf{y}_1) < \hat{f}(\mathbf{y}^{\text{opt}}, \mathbf{y}_2) - \zeta) \rightarrow (f(\mathbf{g}, \mathbf{y}_1) < f(\mathbf{g}, \mathbf{y}_2)), \quad (6)$$

where \mathbf{y}_1 and \mathbf{y}_2 are two Boolean vectors. Note that, for $\zeta = 0$, Equation (6) reduces to the requirement for strict monotonicity, and that for larger $\zeta > 0$ this requirement is increasingly relaxed. Even though this seems trivial, the value of this relation becomes clear by considering that if there exists a strong positive correlation between \hat{f} and f , there may in fact exist a small ζ for which Equation (6) is true.

Based on Lemma 1 and Equation (6), we observe that solutions that are suboptimal in terms of \hat{f} may still be optimal in terms of f , since ζ can be non-zero. In the following, let $\{\mathbf{y}_i; i=1, 2, \dots\}$ and $\{C_i; i=1, 2, \dots\}$ be all the network outputs and the sample sets for the solutions for which holds that $\hat{f}^* - \zeta \leq \hat{f}(\mathbf{y}^{\text{opt}}, \mathbf{y}_i) \leq \hat{f}^*$, respectively. Finally, let ζ be chosen such that Equation (6) holds. Our main theorem can now be formulated as follows:

THEOREM 2.

$$\mathbf{n}^* \in \bigcup_{\forall C_i} V^{(C_i)} \quad (7)$$

PROOF. First, assume that Equation (6) holds for $\zeta = 0$, and thus $\hat{f}(\mathbf{y}^{\text{opt}}, \mathbf{y}_i) = \hat{f}^* \forall i$. Furthermore, from Equation (6) it follows that in this case there exists a direct one-to-one relation between \hat{f} and f . Consequently, a maximum in \hat{f} is guaranteed to correspond to a maximum in f and $V^{(C_i)}$ must contain \mathbf{n}^* . This is true because from Lemma 1 it follows that $V^{(C_i)} \neq \emptyset$. For non-zero values of ζ , the one-to-one relation does not hold. However, from Equation (6), it follows that all values of f for which the corresponding \hat{f} lies outside the interval $[\hat{f}^* - \zeta, \hat{f}^*]$ are strictly smaller than the value of f corresponding to \hat{f}^* . Thus, it must be the case that the maximum of f is constrained to solutions for which \hat{f} lies in the interval $[\hat{f}^* - \zeta, \hat{f}^*]$. Therefore, the union of the sets of solutions that lie in this interval will contain \mathbf{n}^* . ■

From Theorem 2 it naturally follows that:

COROLLARY 3.

$$\mathbf{n}^* = \underset{\mathbf{n}}{\operatorname{argmax}} f(\mathbf{g}, \mathcal{L}(\mathbf{L}, \mathcal{B}, \mathbf{n})) \forall \mathbf{n} \in V^{(Q)}, \quad (8)$$

where $V^{(Q)} = \bigcup_{\forall C_i} V^{(C_i)}$. Notably, if there exists a small ζ for which Equation (6) holds, the number of solutions in $V^{(Q)}$ is limited, and hence \mathbf{n}^* is easily determined by an exhaustive search over all possible solutions in $V^{(Q)}$. In the following, we show that in practice the set $V^{(Q)}$ is small by choosing \mathbf{w} such that ζ is small.

Estimating the weights

Ideally, vector \mathbf{w} is chosen such that ζ is minimal. For practical purposes, it is sufficient to choose \mathbf{w} so that ζ is small, which can be realized by minimizing the difference between \hat{f} and f . For this purpose, we sample the (\hat{f}, f) -relation

by generating N random instances \mathbf{y}_n by introducing up to m random bit-flips in \mathbf{y}^{opt} (shown in Figures 2B and C). The N corresponding association measures f_n and Hamming similarities are collected in vector $\mathbf{f}=[f(\mathbf{g},\mathbf{y}_1),f(\mathbf{g},\mathbf{y}_2),\dots]^T$ and matrix $\hat{\mathbf{F}}=[(\mathbf{y}^{\text{opt}}\leftrightarrow\mathbf{y}_1)^T,(\mathbf{y}^{\text{opt}}\leftrightarrow\mathbf{y}_2)^T,\dots]^T$, respectively. In the latter, \leftrightarrow denotes the XNOR operation, which evaluates to '1' in case its arguments are equal. Notably, m (the number of bit-flips) should be chosen such that the region of interest of the distribution of f is sampled. Since we are interested only in network outputs that associate well with the gene expression, we can choose m rather small to focus only on the right tail for which a good fit between \hat{f} and f is obtained. We found that smaller residuals were obtained by converting log-transformed f -values to z -scores, i.e. $\tilde{\mathbf{f}}=z(\ln\mathbf{f})$. Furthermore, to deal with the intercept, the matrix \mathbf{F} is mean centered, denoted by $\hat{\mathbf{F}}$. Using the vector $\tilde{\mathbf{f}}$ and matrix $\hat{\mathbf{F}}$ we can find the weights \mathbf{w} by constraint linear least squares minimization:

$$\mathbf{w}=\underset{\mathbf{w}}{\text{argmin}}\|\tilde{\mathbf{f}}-\hat{\mathbf{F}}\mathbf{w}\|_2, \quad \text{subject to: } w(\tau)\geq w_\varepsilon \quad (9)$$

where $w_\varepsilon > 0$ is a small scalar that ensures each sample receives a non-zero weight. Figure 2 illustrates a typical example showing that the trend of the relation is monotonically increasing, and the spread around the trend is marginal, indicating that Equation (6) indeed holds for a small ζ .

Estimating ζ

The parameter ζ can be estimated by randomly resampling the (\hat{f},f) -relation using the obtained weights and measuring the spread around the trend in the data in the \hat{f} direction (Figure 2C illustrates this schematically). To this end, lowess smoothing was performed to obtain the trend in the data.¹⁸ Subsequently, the spread around this trend was obtained by applying a sliding window in the \hat{f} direction and defining ζ as the maximum spread across all window positions.

Branch and bound search tree

Equation (5) naturally fits a branch and bound framework with a backtracking search tree in which each node corresponds to a particular set of samples C (shown in Figure S2). Although this tree exhaustively represents all possible sample sets C , the search is very efficient since most nodes can be pruned from the search tree. First of all, if $V^{(C)}$ becomes equal to the empty set, all child nodes of node C can be discarded because these will also result in the empty set. Secondly, as a result of the search tree topology, for each node C we can define an upper bound $\hat{f}_{\text{up}}^{(C)}$ and lower bound $\hat{f}_{\text{low}}^{(C)}$. The upper bound $\hat{f}_{\text{up}}^{(C)}$ is defined as the value of \hat{f} that would be obtained assuming all its subnodes do not result in the empty set (best case scenario):

$$\hat{f}_{\text{up}}^{(C)}=\sum_{\tau\in C}w(\tau)+\sum_{\tau\in C_{\text{sub}}}w(\tau), \quad (10)$$

where C_{sub} denotes the collection of all samples in the subnodes of C . The lower bound $\hat{f}_{\text{low}}^{(C)}$ is defined as the value of \hat{f} that would be obtained assuming all subnodes will result in the empty set (worst-case scenario):

$$\hat{f}_{\text{low}}^{(C)} = \sum_{\tau \in C} w(\tau). \quad (11)$$

A vast reduction of the search space is realized by considering the following branch and bound principle: any node C_α can be pruned if there exists a node C_β , for which the following is true:

$$\hat{f}_{\text{up}}^{(C_\alpha)} < \hat{f}_{\text{low}}^{(C_\beta)} - \zeta, \quad \text{under the condition: } V^{(C_\beta)} \neq \emptyset \quad (12)$$

Thus, if we encountered a branch whose worst-case error is better than the best-case error of another branch, we can safely discard the latter.

After the complete search tree is traversed, the set $V^{(\mathcal{Q})}$ is determined by the union of all the nodes that resulted in a non-empty $V^{(C)}$. In Equation (12), the parameter ζ is included to ensure that set $V^{(\mathcal{Q})}$ includes \mathbf{n}^* (Theorem 2). An optimal leaf ordering is obtained when the samples are sorted based on their weight $w(\tau)$. This ensures that $\hat{f}_{\text{up}}^{(C)}$ decreases as quickly as possible, in effect pruning the tree early in the search. Also note that most $V^{(C)}$ will contain many duplicates when symmetries in the topology of \mathcal{B} are considered. By filtering these from $V^{(C)}$ before evaluating the succeeding node results in an additional search speed-up.

Tolerance level

A final, yet influential, search-space reduction is achieved by only considering solutions for which a certain minimum level of association is achieved. This is realized by enforcing that \hat{f}_{low} can never be below a user defined tolerance level. In other words, for this bounded \hat{f}_{low} , we can write: $\hat{f}'_{\text{low}} = \max(\hat{f}_{\text{tol}}, \hat{f}_{\text{low}})$. As a result, branches for which $\hat{f}_{\text{low}} \leq \hat{f}_{\text{tol}}$ can be pruned even before the search is started. The search procedure is explained by example in Figure S2.

Estimating the False Discovery Rate

Because our primary interest lies with the interpretation of the selected genotype markers and combinatorial logic, it is of critical importance to assess frequency of false positives amongst the networks called significant. Due to the efficiency of the proposed method, it is possible to employ a permutation procedure to obtain a null-distribution for each \mathcal{B}_j . From this distribution it is possible to estimate the False Discovery Rate and the associated q -values by using the method proposed by Storey and Tibshirani.¹⁹ Not surprisingly, in many cases, multiple network topologies yield significant associations with the same gene. The q -values, available for each of the solutions, provide a convenient way of performing selection of the most parsimonious model by accepting only the topology for which the q -value is minimal.

URL

The MATLAB code of the prototype implementation is available on: <http://bioinformatics.tudelft.nl/> or <http://bioinformatics.nki.nl/>.

RESULTS

Genetical genomics dataset

The genetical genomics dataset used to demonstrate our method contains genome-wide RNA transcript measurements performed on four related hematopoietic cell populations.¹⁷ These were isolated from the bone marrow of ~25 BXD recombinant inbred mouse strains that were derived by crossing C57BL/6J (B6) and DBA/2J (D2).²⁰ A typical analysis of these data includes determining eQTLs, i.e. regions in the genome for which the genotype across strains associates well with RNA transcript levels.

We inferred associations only for the myeloid cell population, as for this cell type data for the largest number ($T=24$) of unique BXD strains were available. The expression data were pre-processed as described in the Supplementary Methods. Because the CAL networks inferred for highly correlated genes are equivalent, rather than starting the optimization for each gene separately, we constructed gene clusters and searched for CAL networks for the centroids of each gene cluster. To ensure only tightly correlated probes were clustered, we employed a stringent cut-off (correlation distance cut-off 0.2). This resulted in 6139 clusters that were used to determine eQTLs.

Genotype information for the strains was retrieved from The GeneNetworkⁱⁱⁱ. Genotype markers that were highly similar across strains and on the same chromosome were also grouped into clusters to prevent the algorithm from finding many combinations of genotype markers that are equivalent (such as the markers in linkage disequilibrium). This resulted in 453 marker clusters ($L=453$). The cluster centroids were defined as the majority vote of the individual markers in the cluster and were used as putative inputs to the network (see also the Supplementary Methods and Figures S3 and S4).

For setting the tolerance level f_{tol} no straightforward method exists. Preferably, the tolerance level is set close to the final significance threshold to minimize the effort spent on finding optima for gene clusters that can never be significant. We settled for a tolerance level equal to the 75th percentile of the f^{opt} distribution ($f_{\text{tol}}=7.6$), obtained by computing the f -values associated with each \mathbf{y}^{opt} . Gene clusters for which the maximum f -score is below this tolerance level (i.e. in case $f^{\text{opt}} < f_{\text{tol}}$) were not included in the CAL network search, to result in a set of 1525 high-potential gene clusters.

ⁱⁱⁱ <http://www.genenetwork.org/dbdoc/BXDGeno.html> (Accessed: 02 April 2009)

Algorithm performance

From the methods section it follows that, under the condition that an appropriate value for ζ is found, our algorithm produces an optimal solution. We empirically validate this claim by comparing solutions of the proposed algorithm with the global optimum obtained with an exhaustive search. To ensure realistic conditions we do this using the real data described above.

For each gene expression vector we performed our CAL network search as described with seven network topologies containing AND, OR and XOR logic as well as a more complex combination of these Boolean functions. A rather low tolerance level ($f_{\text{tol}}=4$) was used, which turned out to capture most of the solution-space (>80% for all topologies). The solutions obtained were compared with the optimal solutions determined by means of an exhaustive search for the same seven Boolean logic functions using GRID computing facilities. The accuracy is expressed as the percentage of times that the algorithm finds the same solution as the exhaustive search.

Figure 4A shows the resulting accuracy. We observe that for solutions with f -scores between 5 and 6 already >95% accuracy is achieved, while for solutions with f -scores of 6 and higher, virtually 100% accuracy is achieved for each topology. For comparison, Figure 4A also gives the 75th percentiles of the solution distributions for each topology. Because solutions of interest (putatively significant solutions) are required to have f -scores substantially higher than the 75th percentile, we can conclude that our method achieves 100% accuracy for a reasonable operating range (solutions with f -scores between 4 and 5 – where the accuracy is below 95% – are well below the 75th percentile for all networks).

While comparing our method to the method presented in Mukherjee *et al.*,¹² using simulated gene expression vectors and a predetermined random network (ground truth), we found that our method reaches higher true positive rates (see Supplementary Results). These results illustrate the benefit of searching for solutions for each of the network topologies separately, and employing a significance estimate to enforce parsimony.

Obtaining the same accuracy as an exhaustive search is only useful if this is achieved for runtimes that are substantially lower. To assess this, we randomly selected 200 gene expression vectors from the 1525 gene clusters and measured runtimes for both our CAL network search as well as the exhaustive search. Figures 4B-D show these runtimes for a range of conditions. The boxplots represent the results obtained with the CAL network search and the horizontal lines the runtimes for the exhaustive search.

Figure 4B compares runtimes for different network topologies. Clearly, the branch and bound algorithm significantly outperforms the exhaustive search under all experimental conditions with differences in runtime of up to four orders of magnitude. For the three input networks in particular, the runtime required for exhaustive search (more than five hours per gene per network) prohibits any

further permutation procedures. The CAL network search, on the other hand, is able to find the solution in a matter of seconds, thereby enabling the large number of permutations required to obtain reliable significance estimates.

Compared to the variance in runtime of the exhaustive search, which was negligible, the variance of the CAL network search is quite high. This is expected as our CAL network search finishes rapidly when a good solution presents itself early in the search, while more time is needed to conclude that no acceptable solution is present. For a similar reason, the more complex networks, those containing XOR logic, have higher median runtimes. On no occasion, however, does this increase runtimes above 100 seconds for any of the networks.

To evaluate performance as a function for dataset size we artificially increased the number of predictors and the number of samples (Figure 4C). In addition, runtimes for different tolerance levels were examined (Figure 4D). The number of predictors was increased by horizontally concatenating the original matrix L with copies of L containing 10% random bit-flips. The sample size was increased by vertically concatenating matrix L as well as all gene expression vectors \mathbf{g} with copies of L and \mathbf{g} , respectively. In case of the latter, normally distributed noise was added to the copies with $\sigma_{\text{noise}} = 0.1\sigma_{\mathbf{g}}$. We observe that for both the exhaustive search as well as the CAL network search runtimes increase substantially as the number of predictors increases. In case of the CAL network search this is explained by the fact that many very good solutions are present due to the increased imbalance between the number of predictors and the sample size. It is expected, yet not quantitatively established, that better performance is observed when this balance is restored. The increase in runtime as a result of an increased number of samples is moderate, with a median runtime considerably lower than an exhaustive search for only two input networks. Likewise, increasing the tolerance level only moderately speeds up the CAL network search, demonstrating that runtime is robust for the setting of this parameter.

Combinatorial expression Quantitative Trait Loci

We performed the CAL network search for the set of 1525 high potential gene clusters. The complete search (e.g. for all gene clusters and all topologies) was repeated 100 times using a permuted version of the gene-expression vectors. For each topology, this resulted in a null-distribution containing 152500 values, which was used to estimate q -values for each of the resulting solutions. We considered network topologies with a maximum of three inputs listed in Figure S5. Notably, we included two single-input networks to account for direct positive and negative association, respectively, which is equivalent to positive association with the D2 and B6 allele, respectively. This ensures that the algorithm has the option of choosing the least complex model in case an eQTL is capable of explaining a significant portion of the variance in the expression of the gene cluster.

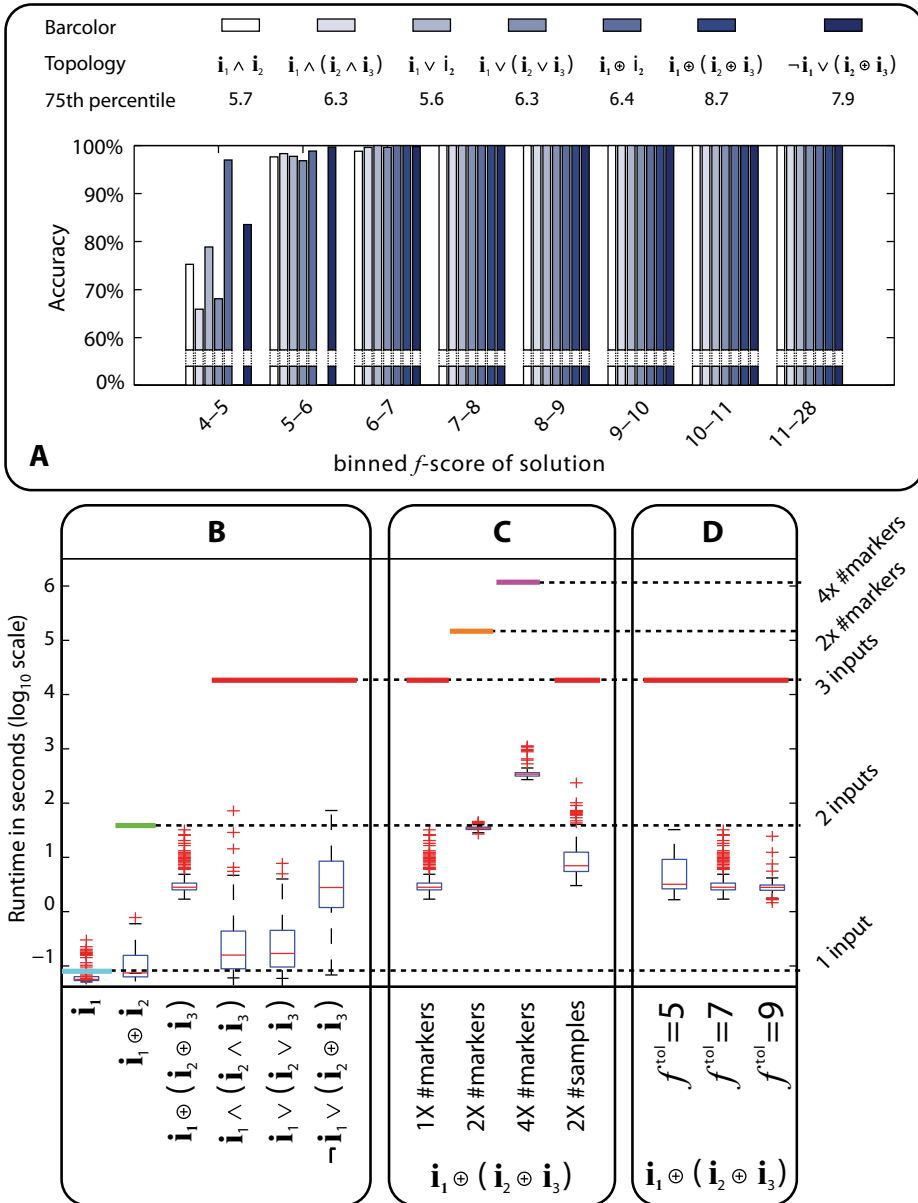


Figure 4. Algorithm performance in terms of accuracy and runtime under various conditions. (A) Bar graph displaying accuracy for different network topologies and different values of the f -score. For each of the network topologies the 75th percentile of the solution distribution is also given, showing that for solutions in the tail 100% accuracy is obtained. For the two missing network topologies, dataset sizes and tolerance levels. (B-D) Runtimes for different network topologies, dataset sizes and tolerance levels. The horizontal lines reflect runtimes for exhaustive search. From bottom to top these represent the runtimes for: a single input network, two input network and three input network with one, two and four times the number of predictors, respectively.

Figure 5A gives an overview of the number of gene clusters for which the output of a CAL network significantly (at the 10% FDR level) associated with its expression (red bars). To obtain additional confidence in the significance threshold we calculated q -values for ten additional permutations of the whole dataset. For none of the network topologies did the mean number of significant gene clusters across the ten permutations exceed 0.6, indicating that the expected number of false discoveries is conservatively kept under control. The yellow bars indicate the number of significant gene clusters after model size selection based on the q -value as detailed in the Methods section. It appears that most of the gene clusters for which association is observed can be explained by one of the single input networks. For nine gene clusters (corresponding to 17 genes), however, a CAL network was capable of explaining significantly more of the variance than one of the single input networks or any one of the other CAL networks.

The network topologies, q -values and association scores of the significant CAL networks are given in Figure 5B. Not surprisingly, for all gene clusters at the output of these networks, the *combination* of loci is vastly superior in explaining the variance in expression over any of the markers in isolation. Interestingly, many of these genomic regions would have been missed, as in seven of the networks the best markers do not coincide with one of the inputs of the CAL network.

The sets of markers that were found as the optimal inputs for the seven topologies were mapped onto the genome. Combinatorial eQTLs (ceQTL) were then defined as stretches of consecutive markers. A genome map of the (c)eQTLs is given in Figure 5C, showing the eQTLs (red and blue crosses for positive and negative association, respectively) and ceQTLs (colored symbols) on the x-axis versus the genomic positions of the probes measuring expression on the y-axis. The numbers near the ceQTL symbols correspond to the inputs of the CAL networks depicted in Figure 5B.

Before we zoom in on one of the CAL networks in more detail, some general observations can be made. In particular, we note that in some cases overlap exists among the markers selected at the inputs of the CAL networks and between other network inputs and eQTLs. In seven instances, the identified ceQTLs coincide with eQTLs (connected by black dashed lines in the figure). Some of these eQTLs are located in *cis*. The finding of CAL networks that share one of their inputs (ceQTLs) with an eQTL suggests that the local genotype associated with the eQTL is involved in the regulation of a local gene (*cis*-regulation), but in addition collaborates with the other CAL input locus/loci to regulate the CAL network output gene(s). Furthermore, two of the CAL networks (ranked sixth and ninth) share a ceQTL between the inputs (connected by red dashed lines). It is not inconceivable that a gene present in this ceQTL is indeed involved in the regulation of the target genes of both networks, but that the interaction partners through which this regulation is established differs for both target genes.

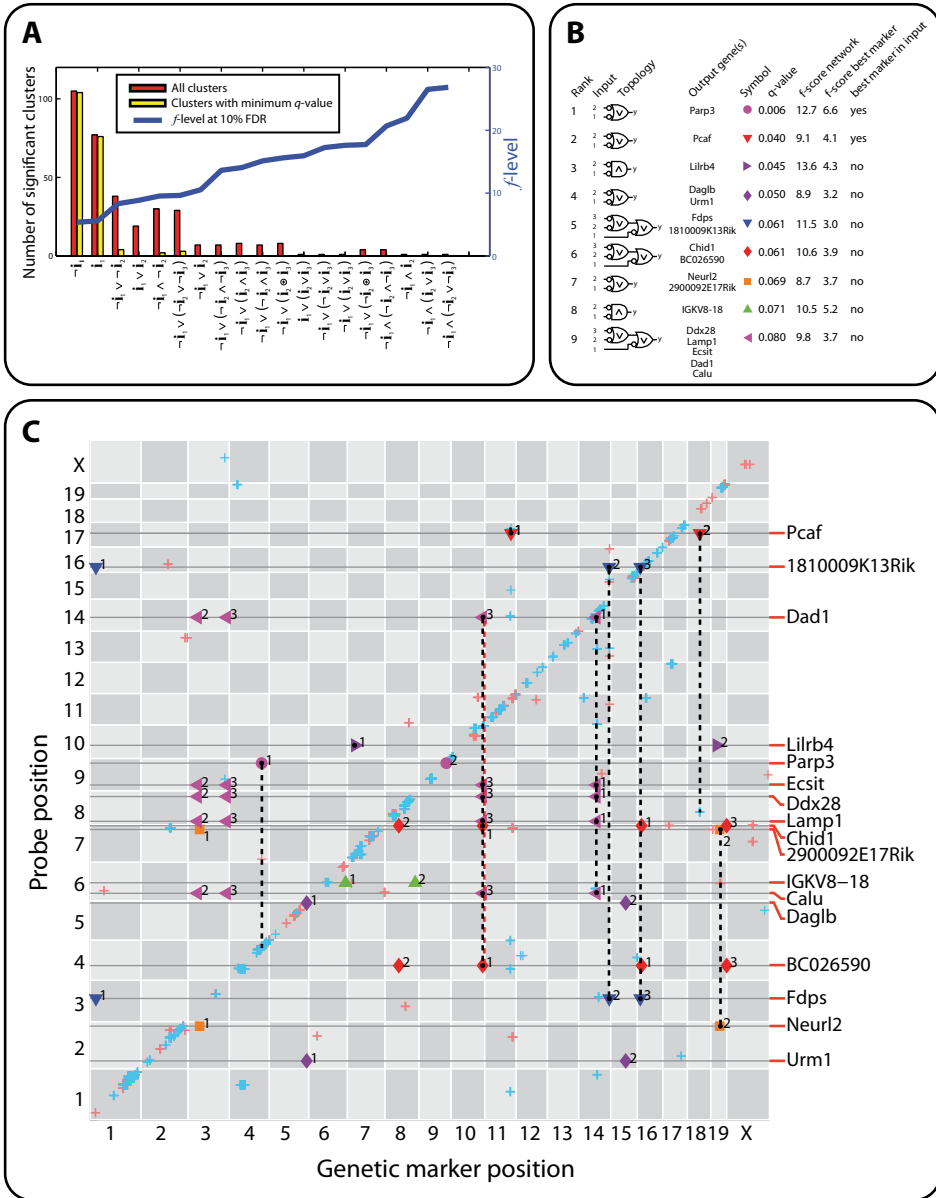


Figure 5. Significant CAL networks. (A) Bar graph with an overview of the number of gene clusters for which a significant (10% FDR) solution is found. Network topologies are sorted according to the 10% FDR level (blue line). (B) CAL networks significant at 10% FDR. The color and shape of the symbols correspond to the symbols used in (C). Small circles at the inputs of the networks denote negation, that is, for these inputs the mapping from allele to binary representation is switched. We also indicate whether the best single marker coincides, for that gene cluster, with one of the inputs of the CAL network. (C) Marker/probe-plot for the top CAL networks showing both the eQTLs (blue crosses) and ceQTLs (sets of colored symbols of various shapes). The colors and shapes of the markers refer to ▶

Among the list of output genes of the nine most significant CAL networks is *Lilrb4* (ranked third). *Lilrb4* encodes a leukocyte immunoglobulin-like receptor that is expressed on the surface of mast cells, neutrophils, and macrophages. It plays a key role in counter-regulating the inflammatory response to prevent pathologic excessive inflammation.²¹

Figure 6 shows small regions around the ceQTLs that were selected as inputs for the CAL network of *Lilrb4*. For each region the association was measured between the expression of *Lilrb4* and the individual markers (blue lines). The red lines, on the other hand, give the association score for the network output. Clearly, the association between the logical combination of inputs and the expression of *Lilrb4* is markedly higher than considering any of the markers in isolation. The regions for which the red curves reach their maximum correspond to the ceQTLs.

The Boolean heat map, displayed at the bottom of Figure 6, outlines the genotype of one particular combination of genetic markers in the ceQTLs across the BXD mouse strains. The bottom two rows of this heat map give the optimal network output and predicted output, respectively. For the *Lilrb4* network the optimal network output is exactly recapitulated by the CAL network. For *Lilrb4* elevated expression is exclusively observed in case of B6 alleles in both the ceQTL regions of Chromosomes 7 and 19.

To focus our attention to the most interesting genes in the ceQTLs we performed a literature search using Ingenuity Pathway Analysis (Ingenuity Systems, www.ingenuity.com). Interestingly, we found a substantial number of interactions between genes localized in the ceQTLs and *Lilrb4*. For example, the literature search revealed a link between *Apba1* (located in the ceQTL region on Chr 19) and *Lilrb4*. Both protein products have been described to bind ITGB3.^{22;23} In addition, the search revealed a link between *Psenen* (Chr 7 ceQTL) and *Apba1* (Chr 19 ceQTL). Both protein products have been described to bind PSEN1 and PSEN2.^{24;25}

While literature is able to link the genes in the ceQTLs to *Lilrb4* and thereby gives the first clues as to how the expression of *Lilrb4* may be regulated, we do not exclude that other interactions (not yet represented in literature) exist. In any case, the result of our method should provide a set of testable hypotheses that can be validated in the laboratory.

- the network topologies listed in (B). Horizontal grey lines connect the inputs and the output of the CAL network. Because probes were clustered, it occurs that the ceQTLs map to multiple probes in case these probes were part of the same cluster. The numeric labels near the colored symbols correspond to the input of the network. Notably, some probes seem to be predicted by more ceQTLs than there are inputs to the CAL network reported. This occurs when there are multiple combinations of markers that show the same association with the gene expression level of the network output, and can be explained by similarity among markers. The cis-band (diagonal) is clearly visible, and in one occasion contains a ceQTL. Overlap among ceQTLs from different networks is marked by red dashed lines, overlap between ceQTLs and eQTLs by black dashed lines.

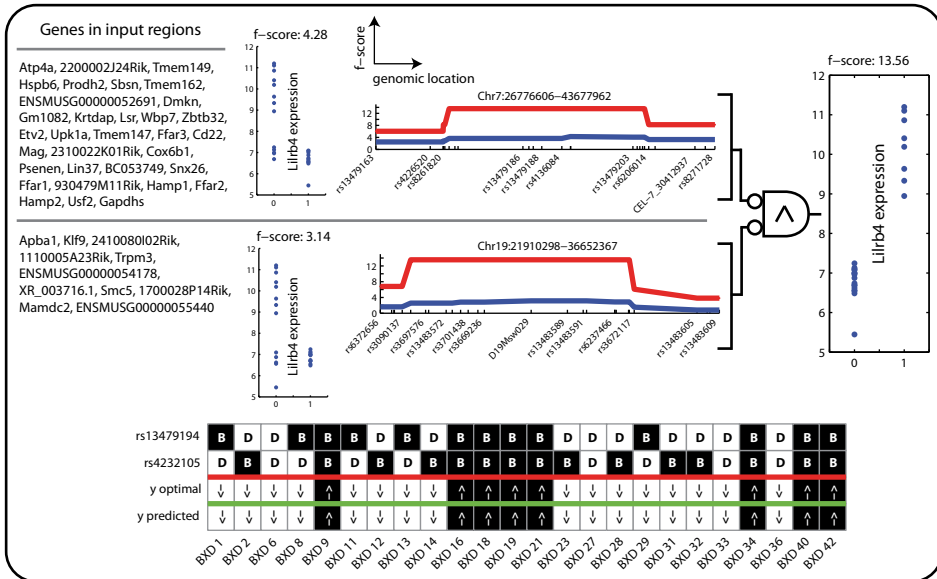


Figure 6. Input regions of the CAL network for *Lilrb4*. The line graphs give the f -score for association between the output gene and the individual markers (blue) and the network output (red). The latter was computed by taking the maximum f -score of the network using the marker under evaluation for one input and any of the other markers for the second input of the network. Where possible the IDs of the genetic markers are given, but some were omitted for readability. The dot plots give the expression values separated by network output (right) and the best markers in the inputs (left). Finally, for one particular combination of markers the genotype for all strains is depicted as a Boolean heat map. In these diagrams, the NOT gates were already incorporated.

DISCUSSION

Unraveling (transcriptional) regulatory networks by inferring complex associations, for instance between genotype and gene expression, necessitates algorithms that take into account possible (allele-specific) interactions. For this purpose, we have proposed a method to efficiently infer CAL networks, i.e. small logic networks in which allele-specific interactions are modeled by Boolean functions. To find the best possible fit of the model given the data, a computationally challenging optimization problem had to be solved. This was achieved by rewriting the optimization such that it could be effectively solved by a customized branch and bound algorithm. Proof and empirical evidence for optimality of the solution, under appropriate conditions, was given. At the same time, differences in runtimes of up to four orders of magnitude were observed when compared to exhaustive search.

Because the CAL network search is able to find the optimal solution in a matter of seconds a permutation procedure becomes feasible, which can be employed to obtain estimates of the False Discovery Rate. This is a major advantage as

the resulting q -values allow selection of the most parsimonious model and enable ranking the network topologies in terms of their complexity.

We demonstrated our algorithm on a genetical genomics dataset, and found that, from the 1525 gene clusters (2913 genes) that resulted after selection of high potential genes, nine gene clusters (17 genes) were significantly associated (at 10% FDR level) through a logical combination of genomic loci rather than a single eQTL. Notably, without incorporating the complex interactions, these associations would have gone unnoticed. Many of the discovered input regions were found to overlap eQTLs or were shared inputs of CAL networks explaining the expression of other genes, suggesting that these regions, indeed, are involved in transcriptional regulation.

ACKNOWLEDGMENTS

We thank Daoud Sie and Leonid V. Bystrykh for their valuable input. This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), and has partially been supported by the Netherlands Genomics Initiative (Horizon, 050-71-055); the Dutch e-Science Grid BiG Grid, SARA - High Performance Computing and Visualisation; the Dutch Cancer Society (RUG2007-3729); the Netherlands Organization for Scientific Research (VICI, 918-76-601 to G.d.H.); and by the European Community (EuroSystem, 200720).

SUPPORTING INFORMATION AVAILABLE ONLINE

Supplementary Methods

- Computation of matrix \tilde{V}
- Branch and bound search tree
- Pre-processing
- CAL network search
- Post-processing

Supplementary Results

- Comparison with Mukherjee *et al.*

Supplementary Figures

- Figure S1. Continued example from Figure 3
- Figure S2. Search tree for the example in the Methods section
- Figure S3. Flow diagram of the pre-processing and CAL network search
- Figure S4. Clustering of genomic markers
- Figure S5. Network topologies evaluated in this study
- Figure S6. Quantitative comparison of CAL network inference with the SCI method

REFERENCES

1. Bystrykh L, Weersing E, Dontje B et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat.Genet.* 2005;37(3):225-232.
2. Pollack JR, Sorlie T, Perou CM et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc.Natl.Acad.Sci.U.S.A* 2002;99(20):12963-12968.
3. Shames DS, Girard L, Gao B et al. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. *PLoS.Med.* 2006;3(12):e486.
4. Visel A, Blow MJ, Li Z et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 2009;457(7231):854-858.
5. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Genet.* 2001;17(7):388-391.
6. Schadt EE, Monks SA, Drake TA et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 2003;422(6929):297-302.
7. Frankel WN, Schork NJ. Who's afraid of epistasis? *Nat.Genet.* 1996;14(4):371-373.
8. Michaelson JJ, Loguercio S, Beyer A. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* 2009;48(3):265-276.
9. Manichaikul A, Moon JY, Sen S, Yandell BS, Broman KW. A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics* 2009;181(3):1077-1086.
10. Wongsee W, Assawamakin A, Piroonratana T et al. Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. *BMC.Bioinformatics.* 2009;10:294.
11. Nunkesser R, Bernholt T, Schwender H, Ickstadt K, Wegener I. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics.* 2007;23(24):3280-3288.
12. Mukherjee S, Pelech S, Neve RM et al. Sparse combinatorial inference with an application in cancer biology. *Bioinformatics.* 2009;25(2):265-271.
13. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nat.Genet.* 2007;39(9):1167-1173.
14. Evans DM, Marchini J, Morris AP, Cardon LR. Two-stage two-locus models in genome-wide association. *PLoS.Genet.* 2006;2(9):e157.
15. Ljungberg K, Holmgren S, Carlborg O. Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. *Bioinformatics.* 2004;20(12):1887-1895.
16. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. *Genet.Epidemiol.* 2005;28(2):157-170.
17. Gerrits A, Li Y, Tesson BM et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS. Genet.* 2009;5(10):e1000692.
18. Cleveland, W. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 1979;74:829-836.
19. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc.Natl. Acad.Sci.U.S.A* 2003;100(16):9440-9445.
20. Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC.Genet.* 2004;5:7.
21. Katz HR. Inhibition of pathologic inflammation by leukocyte Ig-like receptor B4 and related inhibitory receptors. *Immunol.Rev.* 2007;217:222-230.
22. Calderwood DA, Fujioka Y, De Pereda JM et al. Integrin beta cytoplasmic domain interactions with phosphotyrosine-binding domains: a structural prototype for diversity in integrin signaling. *Proc.Natl. Acad.Sci.U.S.A* 2003;100(5):2272-2277.
23. Castells MC, Klickstein LB, Hassani K et al. gp49B1-alpha(v)beta3 interaction inhibits antigen-induced mast cell activation. *Nat. Immunol.* 2001;2(5):436-442.
24. Biederer T, Cao X, Sudhof TC, Liu X. Regulation of APP-dependent transcription complexes by Mint/X11s: differential functions of Mint isoforms. *J.Neurosci.* 2002;22(17):7340-7351.
25. Steiner H, Winkler E, Edbauer D et al. PEN-2 is an integral component of the gamma-secretase complex required for coordinated expression of presenilin and nicastrin. *J.Biol.Chem.* 2002;277(42):39062-39065.

CHAPTER 5

GENETIC SCREEN IDENTIFIES *ZFP521*
AS A CANDIDATE REGULATOR OF
HEMATOPOIETIC STEM CELL POOL SIZE

Alice Gerrits, Yang Li, Bruno M. Tesson,
Rainer Breitling, Berthold Göttgens, Ritsert C. Jansen,
Leonid V. Bystrykh and Gerald de Haan

In preparation

ABSTRACT

Hematopoietic cell differentiation is accompanied by profound changes in gene expression patterns. To identify the regulatory mechanisms that underlie these expression changes, We previously performed a combined analysis of variation in gene expression across four developmentally related hematopoietic cell populations and across mouse strains (C57BL/6, DBA/2, and their recombinant inbred strains). This genetic analysis provided a genome-wide overview of cell-type-specific transcripts and the genomic loci that were most likely to control their expression levels. Here, we illustrate how the presence of genetic diversity in a pedigree of normal, wild-type, mice can be exploited to construct gene networks that operate in successive stages of cellular development. We start with the reconstruction of an experimentally validated gene network, and proceed with the construction of a novel gene network for transcripts that are most specific to hematopoietic stem cells. Finally, we identify *Zfp521* as a strong candidate regulator of hematopoietic stem cell pool size, and uncover the network members through which this factor may operate.

INTRODUCTION

Hematopoietic stem cells (HSCs) are at the apex of the hematopoietic hierarchy. They possess self-renewal activity in order to maintain their original pool, as well as the capacity to differentiate into mature blood cells. Genome-wide gene expression studies of different stages of hematopoietic development have provided insight into the molecular nature of HSC self-renewal, lineage specification, commitment and differentiation.¹⁻⁴ However, the regulatory mechanisms underlying the gene expression changes during hematopoiesis remain largely elusive. The search for these mechanisms has transformed into a virtual “holy grail” of experimental hematology, as the dissection of these mechanisms would not only provide insight into normal and malignant hematopoiesis, but would also pave the way for effective HSC expansion protocols.

The size of the HSC pool is subject to extensive variation among inbred mouse strains and a genetic basis for much of this variation is now well established. Specifically, C57BL/6 (B6) and DBA/2 (D2) mouse strains show a remarkable difference in HSC numbers.⁵⁻⁷ BXD (offspring of B6 and D2) recombinant inbred mouse strains have proven to be a useful panel in which to identify the genetic determinants underlying this variation. Using a quantitative trait locus (QTL) mapping approach multiple genomic loci have been identified that regulate HSC pool size.⁷⁻⁹ In order to identify the quantitative trait genes (QTGs) underlying this trait, and to study the genetic basis of variation in gene expression, we previously performed an expression QTL (eQTL) study in four distinct, but hierarchically related hematopoietic cell populations isolated from the BXD panel.¹⁰

In the current study, we exploit the naturally occurring genetic diversity in the BXD panel to construct gene networks. We show that a combined analysis of variation in gene expression across cell types and variation across mouse strains provides a platform for gene network construction. Also, we show that it allows us to evaluate the complex dynamics of gene regulatory relationships during differentiation. We validate our approach by reconstructing an experimentally verified gene network, and continue with constructing a novel gene network for transcripts that are specific to HSCs. Finally, a combinatorial analysis of classical QTLs and eQTLs points to a zinc-finger transcription factor as a strong candidate regulator of HSC pool size.

METHODS

The data presented in this report are based on previously published microarray and eQTL mapping data.¹⁰ Briefly, in that study $\text{Lin}^- \text{Sca-1}^+ \text{c-Kit}^+$ cells, $\text{Lin}^- \text{Sca-1}^- \text{c-Kit}^+$ cells, TER-119^+ cells and Gr-1^+ cells were isolated from the bone marrow of ~25 BXD recombinant inbred mouse strains. Total RNA was isolated, labeled and hybridized to Sentrix Mouse-6 BeadChips (Illumina). The raw expression data from all four cell

types were first log₂ transformed and then quantile normalized as a single group. All raw data were deposited in the NCBI Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>, accession number GSE18067). All processed data were deposited in the GeneNetwork (<http://www.genenetwork.org>).¹¹

Mapping of eQTLs

As previously described,¹⁰ the expression data of the four cell types were corrected for batch effect and then analyzed separately by the following ANOVA model:

$$y_i = \mu + Q_i + e_i$$

where y_i is the gene's log intensity on the i th microarray; μ is the mean; Q_i is the genotype effect under study; and e_i is the residual error.

For each probe, we performed a genome-wide linkage analysis and identified the marker that showed the most significant eQTL. We defined an eQTL as *local* if it was located within less than 10 Mb from the gene. All other eQTLs were considered *distant*.

Extraction of shared eQTLs

We used a relaxed threshold of $-\log_{10} p > 3$ and retrieved all the markers within the eQTL confidence intervals defined with a 1.5 eQTL drop-off value. When two transcripts had overlapping confidence intervals, they were considered to be co-mapping.

RESULTS

Dynamics of gene expression

We evaluated genome-wide gene expression levels in Lin⁻Sca-1⁺c-Kit⁺ multilineage cells, committed Lin⁻Sca-1⁻c-Kit⁺ cells, TER-119⁺ erythroid cells and Gr-1⁺ myeloid cells isolated from a large panel of BXD recombinant inbred mouse strains. Since we profiled multiple developmentally related cell stages, we were able to illustrate the dynamics of gene expression along either the erythroid (Figure 1A) or myeloid differentiation trajectory (Figure 1B). The 120 transcripts (represented by 138 probes) that were consistently up-regulated during erythropoiesis included well-known erythroid-specific genes such as *Epor*, *Klf1*, *Kel*, *Ank1*, *Gypa*, *Tfrc*, and *Gata1*, and the 29 transcripts (represented by 29 probes) that were consistently up-regulated during myelopoiesis included *Cebpe*, *Ccnd3*, *Idh1*, and *Fcgr3*. We identified 163 transcripts (174 probes) that were repressed during erythroid development, and 88 transcripts (92 probes) that were repressed during myeloid development. Only 52 transcripts were consistently repressed in both differentiation trajectories. These transcripts included the well-known stem cell

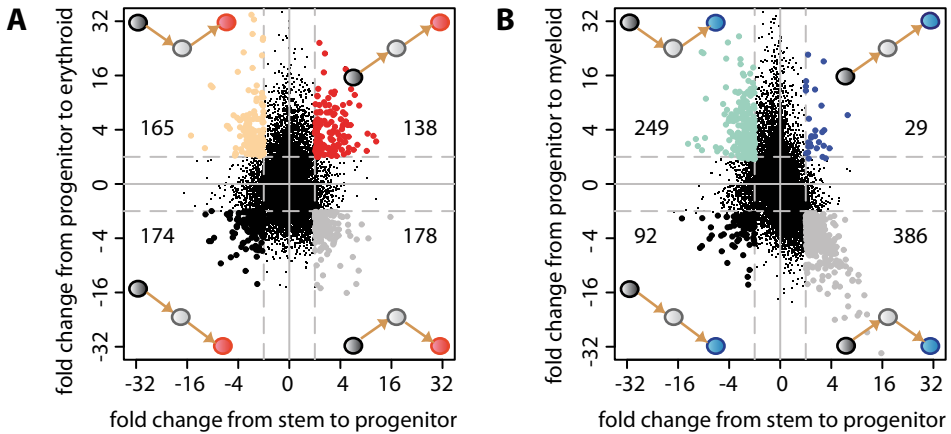


Figure 1. Mean expression levels for all probes in the four cell types. (A) Gene expression trajectory during erythroid differentiation. Gene expression patterns are shown as cells differentiate from $\text{Lin}^{-}\text{Sca-1}^{+}\text{c-Kit}^{+}$ to $\text{Lin}^{-}\text{Sca-1}^{-}\text{c-Kit}^{+}$ to TER-119^{+} cells. Each dot refers to the mean intensity of a single probe across all BXD strains. Dotted lines indicate 2-fold change thresholds. The bold and colored dots refer to those transcripts that met the criteria of the respective quadrants. For example, the 138 transcripts for which expression increased 2-fold as cells differentiated from $\text{Lin}^{-}\text{Sca-1}^{+}\text{c-Kit}^{+}$ to $\text{Lin}^{-}\text{Sca-1}^{-}\text{c-Kit}^{+}$ (x-axis), and again 2-fold from $\text{Lin}^{-}\text{Sca-1}^{-}\text{c-Kit}^{+}$ to TER-119^{+} (y-axis), are shown in red. (B) As in panel A, but here mean gene expression levels during myeloid development to Gr-1^{+} cells are shown.

genes *Meis1*, *Hoxa5*, *Hoxa7*, *Dnmt3b*, and *Cd34*, but also many poorly annotated transcripts. Genes belonging to each of the different gene expression quadrants are provided in Table S1.

Reconstruction of an experimentally validated network

Unraveling the gene regulatory networks that control important cell fate decisions has become a focal point in the field of hematology. Such networks are often visualized as networks composed of nodes (representing genes, proteins and/or metabolites) and edges (representing molecular interactions). In the hematopoietic system, interactions between several key transcription factors have been experimentally validated. Antagonistic interactions between two key regulators, namely the *Gata1* and *Pu.1* transcription factors, are thought to control the erythroid versus myeloid fate choice in common myeloid progenitors as indicated by both gain- and loss-of-function studies in cell line systems.^{12,13} Moreover, experiments in transgenic mice as well as cell line model systems suggested that the key HSC regulators *Tal1* (also known as *Scf*) and *Gata2* together with Ets-family transcription factors such as *Pu.1* and *Fli1* can form a triad network motif characterized by extensive positive feedback loops which might be important in maintaining the undifferentiated state in primitive hematopoietic cells.¹⁴⁻¹⁶ Since *Gata1* together with *Zfp1* (also known as *Fog1*) is known to repress *Gata2*, it is

likely that there will be cross-talk between these two subcircuits at intermediate stages of hematopoietic differentiation (Figure 2A).¹⁷

Given that the generation of the above network models relied on assays that experimentally perturb the normal equilibrium of regulatory interactions, we reasoned that our expression dataset in genetically diverse, yet wild-type, primary hematopoietic cell types should provide an ideal resource to probe the nature of any proposed interactions in the four cell types sampled in the current study. To identify correlations between pairs of genes within and between the

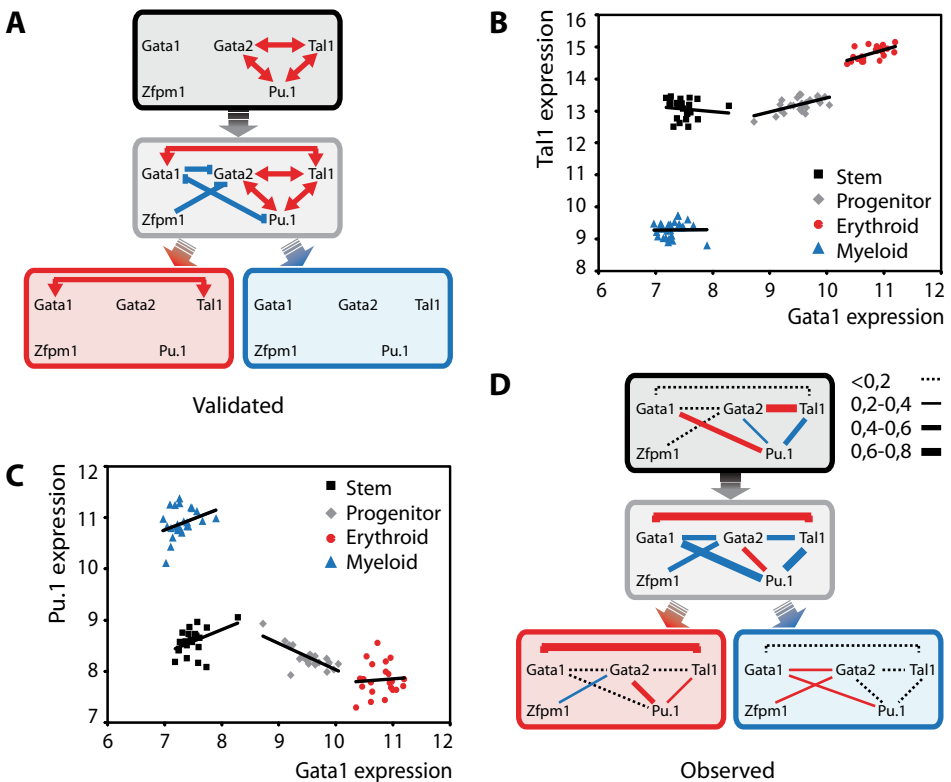


Figure 2. Network dynamics. (A) Experimentally validated transcription factor interactions in each of the four cell types, based on previously published data. $Lin^{-}Sca-1^{+}c-Kit^{+}$ cells are shown in black (top), $Lin^{-}Sca-1^{-}c-Kit^{+}$ cells in grey (middle), $TER-119^{+}$ cells in red (bottom left), and $Gr-1^{+}$ cells in blue (bottom right). (B) Observed correlation between variation in *Gata1* and *Tal1* expression levels across cell types and across mouse strains. In this scatter plot, each dot represents *Gata1* and *Tal1* expression levels for a single BXD strain in a single, indicated, cell type. (C) Correlation between *Gata1* and *Pu.1* expression. (D) Correlation coefficients of all predicted gene pairs in the validated network as shown in panel A. Red and blue lines connect positively and negatively correlated pairs of genes, respectively. Solid lines indicate absolute correlation coefficients larger than 0.2, and the width of the lines reflect the strength of the correlation.

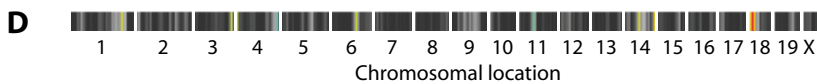
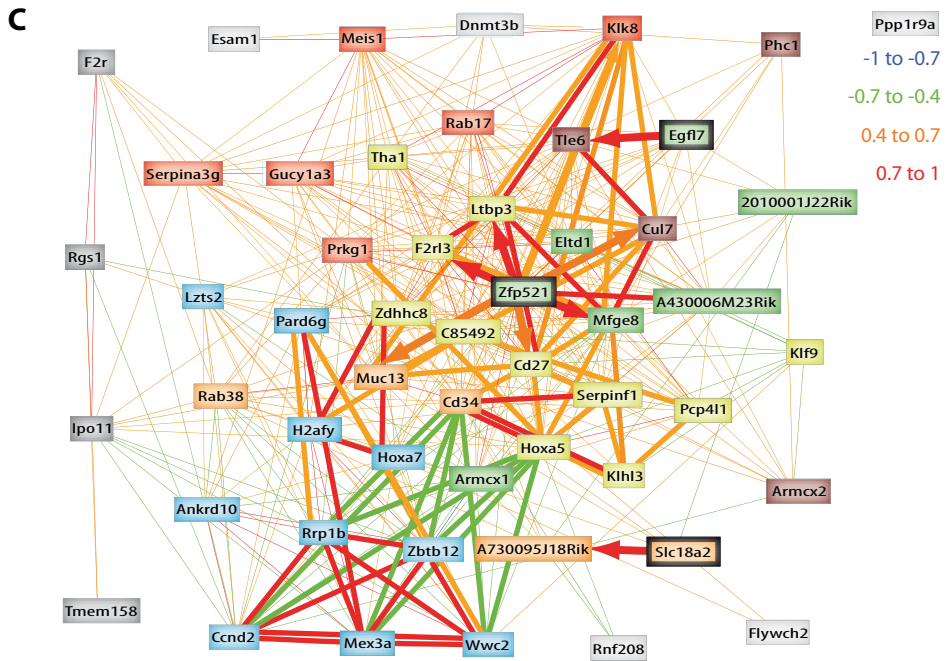
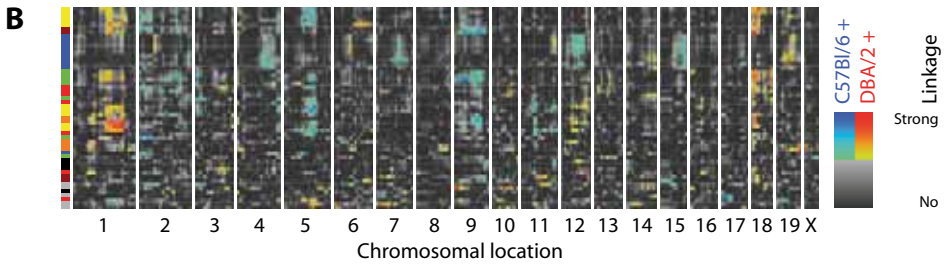
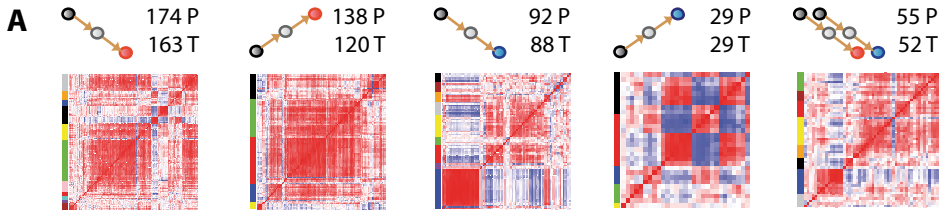
two subcircuits, we plotted gene expression levels across all mouse strains in each of the four cell types (Figures 2B—C), determined the associated linear correlation coefficients, and superimposed these correlation measures on the predicted network model. This analysis confirmed many of the proposed positive and negative interactions, either as positive or as negative correlations (Figure 2D). Remarkably, both the strength and the direction of correlations were highly dynamic across the four cell populations. For example, while the predicted antagonistic relationship between *Gata1* and *Pu.1* in $\text{Lin}^{-}\text{Sca-1}^{-}\text{c-Kit}^{+}$ cells was indeed observed, the corresponding correlations were positive in $\text{Lin}^{-}\text{Sca-1}^{+}\text{c-Kit}^{+}$ and myeloid cells. Our analysis therefore not only revealed *a priori* unexpected dynamics of regulatory interactions but also provided a novel strategy to pinpoint specific cell types along a differentiation cascade where particular interactions may be critical.

Construction of gene networks in hematopoiesis

Having established that the combined analysis of variation in gene expression across cell types and across mouse strains is able to confirm genetic interactions of known key transcription factors, and reveal complex dynamics of gene regulatory relationships during differentiation, we proceeded to expand our analysis to explore novel regulatory networks. To this end, we performed hierarchical clustering of all genes that were consistently up- or down-regulated during erythroid or myeloid development (bottom left and top right quadrants in Figures 1A—B), based on inter-strain correlation of variation in expression in $\text{Lin}^{-}\text{Sca-1}^{-}\text{c-Kit}^{+}$ cells. One could hypothesize that the coordinated up- or down-regulation of collections of transcripts during cellular development results from a limited number of key instructive decisions. If this were the case, strong correlation structures of co-regulated transcripts should exist. Indeed, we detected multiple modules of co-varying transcripts for each set of consistently up- or down-regulated transcripts (Figure 3A).

To reveal whether such modules of transcripts were under common genetic control, we performed eQTL clustering of those 52 transcripts (55 probes representing 52 transcripts) that were consistently down-regulated during both erythroid and myeloid differentiation (Figure 3B). A genetic analysis in $\text{Lin}^{-}\text{Sca-1}^{-}\text{c-Kit}^{+}$ cells revealed co-localization of eQTLs for multiple transcripts, which suggests shared regulatory relationships. Interestingly, a substantial number of the 52 transcripts were regulated by loci on chromosomes 1, 5, 9, 12, and 18. This is consistent with earlier QTL studies which documented that these chromosomes contain loci that control the pool size or the turnover of primitive hematopoietic cells.^{6-8;18;19}

To construct a regulatory network for these 52 transcripts, we used *GeneNetwork* to create an association network based on their pairwise Pearson correlation coefficients (Figure 3C). In order to extract shared eQTLs, for each transcript we not



only identified its maximal eQTL, but all eQTLs above a relaxed threshold of $p < 10^{-3}$. Two transcripts were considered to be co-mapping when they had overlapping eQTL intervals. The results of this analysis were superimposed on the association network (Figure 3C). When one of the transcripts was regulated *locally*, direction of causality could be established, as a *locally* regulated gene is predicted to be upstream of a *distantly* co-regulated gene. Interestingly, many transcripts were transcriptionally controlled by a locus on chromosome 18, containing *Zfp521* (one of the 52 transcripts that were consistently down-regulated during differentiation). Variation in *Zfp521* expression mapped to two loci on chromosome 18, one of which contained the gene encoding *Zfp521* itself. Strikingly, the classical trait HSC pool size had previously been mapped to this exact same locus (Figure 3D), thereby pointing to *Zfp521* as a candidate regulator of this trait.⁹

Zfp521 was discovered in 2003 as a common retroviral integration site (and was termed ecotropic viral integration site 3 or *Evi3*) in murine B-cell lymphoma and encodes a Krüppel-like zinc finger protein.²⁰ Since then, the expression of the human homolog of *Zfp521*, EHZF, has been shown to be restricted to CD34⁺ progenitor cells, and to be expressed in many acute myeloid leukemias.²¹ That *Zfp521* has a functional role in immature cells became apparent when overexpression of EHZF in primary human CD34⁺ cells was described to result in inhibition of differentiation and in the accumulation of progenitors *in vitro*.²² Our data confirm a central role for the transcription factor *Zfp521* in regulating transcripts that are specific to the most primitive hematopoietic cells, and identify candidate network members through which this factor may regulate HSC pool size.

- ◀ **Figure 3. Gene network construction.** (A) Hierarchical clustering of those transcripts that were consistently down- or up-regulated during erythroid and/or myeloid differentiation (as defined in Figures 1B–C). Clustering was based on variation in transcript abundance across BXD strains in Lin⁻Sca-1⁻c-Kit⁺ cells. Positive and negative correlations are shown in red and blue, respectively. Several modules of co-regulated genes (indicated by the colored squares next to the heatmap) were identified. The rightmost panel clusters 52 transcripts (represented by 55 probes) that were down-regulated during both erythroid and myeloid commitment. (B) Genetic analysis of the 52 transcripts shown in panel A identified shared eQTLs. This plot shows a genome-wide heatmap of all 52 transcripts where the color coding indicates strength of the genetic association and variation in expression. Significant eQTLs are indicated with red and blue colors for B6 or D2 alleles increasing expression, respectively. The color coded bar left to the figure refers to the clusters of correlated transcripts in panel A. (C) Association network of the 52 transcripts (shown in panel A) in Lin⁻Sca-1⁻c-Kit⁺ cells. Edges show Pearson correlation coefficients < -0.4 or > 0.4 , their lengths being inversely correlated with the correlation coefficients. In this configuration the total stretching of the edges is minimized. Edge colors indicate strengths and directions of the correlation coefficients. Edges are shown in bold when two transcripts are co-mapping to the same locus. Three transcripts (*Egfl7*, *Slc18a2*, and *Zfp521*) are (at least in part) *locally* regulated (encircled in bold black). In these cases, directionality of regulatory input (“who regulates who”) is indicated with an arrow (pointing away from the potential regulator). Colors of boxes represent the correlation clusters to which the transcripts were assigned in panel A. (D) Genome-wide QTL plot of variation in stem cell pool size shows significant linkage to a locus on chromosome 18 (GeneNetwork Trait ID: 10056).

DISCUSSION

5

To decipher the regulatory mechanisms that underlie the changes in gene expression that accompany hematopoietic cell differentiation, we analyzed gene expression variation in cell types isolated from genetically distinct mouse strains. We show how naturally occurring genetic diversity across mouse strains can be exploited to construct gene regulatory networks. We started with the reconstruction of an experimentally verified gene network and illustrate how hematopoietic differentiation could be the result of cell-type dependent rewiring of transcriptional connections. This rewiring is observed as switches in correlation patterns, and indicates reversal of regulatory polarity due to cell-type-specific transcriptional co-factors. We proceeded with the construction of a novel gene network for transcripts that were most highly expressed in HSCs, and identified the zinc finger transcription factor *Zfp521* as a central player in this network. Strikingly, a subsequent combinatorial analysis of classical QTLs and eQTLs identified *Zfp521* as a strong candidate regulator of HSC pool size.

HSCs are known to be controlled by both intrinsic and extrinsic cues. In this report, we show that HSCs may be intrinsically controlled by *Zfp521*. However, *Zfp521* has also recently been shown to regulate the rate of osteoblast differentiation and bone formation.^{23;24} This finding points to a dual role of *Zfp521* in regulating HSCs: 1) intrinsically and 2) extrinsically via the HSC niche. Based on these findings, it would be of great interest to study the HSC niche in the BXD mouse panel. Analyzing the properties of the niche and genome-wide gene expression in the niche would have the potential to unravel (genetic) interactions between HSCs and their micro-environment.

HSC pool size is controlled by multiple genes and is therefore considered a genetically complex trait. Variation in this trait has been mapped to multiple QTLs on chromosomes 1, 3, 5, 11, 18 and the X chromosome.⁷⁻⁹ The main challenge for dissecting complex traits lies not in the identification of QTLs that underlie these traits, but in the identification of the QTGs that underlie them. Previously, *Lxn* was identified as a QTG located in the Chr. 3 QTL interval.²⁵ Here, we report *Zfp521* as a strong candidate QTG in the Chr. 18 QTL interval. It is likely that future studies will reveal additional factors that contribute to the natural variation in HSC numbers. The identification of such factors would be of substantial significance for the fundamental understanding of normal and malignant hematopoiesis, and for exploiting the full clinical potential of HSCs.

ACKNOWLEDGMENTS

We thank Henk Moes and Geert Mesander for assistance in cell sorting; and Rob W. Williams for assistance using The GeneNetwork. This work was supported by the Netherlands Genomics Initiative (Horizon, 050-71-055); by the Netherlands Genomics Initiative/Netherlands Bioinformatics Centre (Biorange grant SP1.2.3); by the Netherlands Organization for Scientific Research (VICI, 918-76-601 to G.d.H.; and VICI, 865-04-001 to R.C.J.); by the Dutch Cancer Society (RUG2007-3729); and by grants from the European Community (Marie Curie RTN EUrythron, MRTN-CT-2004-005499; and EuroSystem, 200720).

SUPPORTING INFORMATION AVAILABLE ON REQUEST

Table S1. Gene expression categories

REFERENCES

- Ivanova NB, Dimos JT, Schaniel C et al. A stem cell molecular signature. *Science* 2002;298(5593):601-604.
- Chambers SM, Boles NC, Lin KY et al. Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell* 2007;1(5):578-591.
- Kiel MJ, Yilmaz OH, Iwashita T et al. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* 2005;121(7):1109-1121.
- Forsberg EC, Prohaska SS, Katzman S et al. Differential expression of novel potential regulators in hematopoietic stem cells. *PLoS.Genet.* 2005;1(3):e28.
- De Haan G, Nijhof W, Van Zant G. Mouse strain-dependent changes in frequency and proliferation of hematopoietic stem cells during aging: correlation between lifespan and cycling activity. *Blood* 1997;89(5):1543-1550.
- De Haan G, Van Zant G. Intrinsic and extrinsic control of hemopoietic stem cell numbers: mapping of a stem cell gene. *J.Exp.Med.* 1997;186(4):529-536.
- Muller-Sieburg CE, Riblet R. Genetic control of the frequency of hematopoietic stem cells in mice: mapping of a candidate locus to chromosome 1. *J.Exp.Med.* 1996;183(3):1141-1150.
- Geiger H, True JM, De Haan G, Van Zant G. Age- and stage-specific regulation patterns in the hematopoietic stem cell hierarchy. *Blood* 2001;98(10):2966-2972.
- Bystrykh L, Weersing E, Dontje B et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat.Genet.* 2005;37(3):225-232.
- Gerrits A, Li Y, Tesson BM et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS.Genet.* 2009;5(10):e1000692.
- Wang J, Williams RW, Manly KF. WebQTL: web-based complex trait analysis. *Neuroinformatics.* 2003;1(4):299-308.
- Zhang P, Behre G, Pan J et al. Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. *Proc.Natl.Acad.Sci.U.S.A* 1999;96(15):8705-8710.
- Rekhtman N, Radparvar F, Evans T, Skoultschi AI. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. *Genes Dev.* 1999;13(11):1398-1411.
- Pimanda JE, Ottersbach K, Knezevic K et al. Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proc.Natl.Acad.Sci.U.S.A* 2007;104(45):17692-17697.
- Wilson NK, Foster SD, Wang X et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide

- analysis of ten major transcriptional regulators. *Cell Stem Cell* 2010;7(4):532-544.
16. Wilson NK, Miranda-Saavedra D, Kingston S et al. The transcriptional program controlled by the stem cell leukemia gene *Scf/Tal1* during early embryonic hematopoietic development. *Blood* 2009;113(22):5456-5465.
 17. Pal S, Cantor AB, Johnson KD et al. Coregulator-dependent facilitation of chromatin occupancy by GATA-1. *Proc.Natl.Acad.Sci.U.S.A* 2004;101(4):980-985.
 18. Chen J, Astle CM, Harrison DE. Genetic regulation of primitive hematopoietic stem cell senescence. *Exp.Hematol.* 2000;28(4):442-450.
 19. De Haan G, Van Zant G. Genetic analysis of hemopoietic cell cycling in mice suggests its involvement in organismal life span. *FASEB J.* 1999;13(6):707-713.
 20. Warming S, Liu P, Suzuki T et al. Evi3, a common retroviral integration site in murine B-cell lymphoma, encodes an EBF1-related Kruppel-like zinc finger protein. *Blood* 2003;101(5):1934-1940.
 21. Bond HM, Mesuraca M, Carbone E et al. Early hematopoietic zinc finger protein (EHZF), the human homolog to mouse Evi3, is highly expressed in primitive human hematopoietic cells. *Blood* 2004;103(6):2062-2070.
 22. Bond HM, Mesuraca M, Amodio N et al. Early hematopoietic zinc finger protein-zinc finger protein 521: a candidate regulator of diverse immature cells. *Int.J.Biochem.Cell Biol.* 2008;40(5):848-854.
 23. Hesse E, Kiviranta R, Wu M et al. Zinc finger protein 521, a new player in bone formation. *Ann.N.Y.Acad.Sci.* 2010;119232-37.
 24. Wu M, Hesse E, Morvan F et al. Zfp521 antagonizes Runx2, delays osteoblast differentiation in vitro, and promotes bone formation in vivo. *Bone* 2009;44(4):528-536.
 25. Liang Y, Jansen M, Aronow B, Geiger H, Van Zant G. The quantitative trait gene *latexin* influences the size of the hematopoietic stem cell population in mice. *Nat.Genet.* 2007;39(2):178-188.

CHAPTER

6

GENETIC SCREEN IDENTIFIES MICRORNA CLUSTER 99B/LET-7E/125A AS A REGULATOR OF PRIMITIVE HEMATOPOIETIC CELLS

Alice Gerrits, Marta A. Walasek, Sandra Olthof,
Ellen Weersing, Martha Ritsema, Erik Zwart,
Leonid V. Bystrykh and Gerald de Haan

Invited for resubmission to Blood

ABSTRACT

Hematopoietic stem/progenitor cell (HSPC) traits differ between genetically distinct mouse strains. For example, DBA/2 mice have a higher HSPC frequency compared to C57BL/6 mice. We performed a genetic screen for microRNAs that are differentially expressed between LSK, LS⁻K⁺, erythroid and myeloid cells isolated from C57BL/6 and DBA/2 mice. This analysis 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 miR-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 upon differentiation. In addition, these microRNAs were higher expressed in DBA/2 cells compared to C57BL/6 cells, and thus correlated with HSPC frequency. To functionally characterize these microRNAs in HSPCs, we overexpressed the entire miR-cluster in BM cells from 5-FU treated C57BL/6 mice and initiated assays to quantify CFU-GM self-renewal, CAFC frequencies, and LT-HSCs. BM cells overexpressing the miR-cluster showed increased replating capacity in CFU-GM assays and dramatically increased day-35 CAFC activity. Furthermore, mice reconstituted with miR-cluster-overexpressing cells displayed severe hematopoietic phenotypes. Finally, we performed genome-wide gene expression arrays and identified candidate functional targets through which this miR-cluster may modulate HSPC fate.

INTRODUCTION

Hematopoietic stem cells (HSCs) sustain life-long blood production, and are characterized by their remarkable self-renewal capacity and multi-lineage differentiation potential. HSCs give rise to progenitors that differentiate into the various blood cell types. Many characteristics of the hematopoietic stem and progenitor cell (HSPC) compartment display mouse strain-dependent variation. For example, compared to C57BL/6 (B6) mice, DBA/2 (D2) mice have a higher HSPC frequency and cycling rate.¹⁻⁴ The observed natural variation between these regular inbred mouse strains can be exploited to identify potential modulators of such HSPC traits. Useful in this respect has been the generation of BXD (progeny of B6 x D2) recombinant inbred mouse strains.^{5,6} Each of these strains represents a unique, but stable genetic mosaic of B6 and D2 alleles. By making use of these strains, variation in many different HSPC traits has been mapped to genomic regions (reviewed in ⁷). Previously, we analyzed mRNA expression variation in developmentally related cell types isolated from the BXD mouse panel.^{8,9} In those studies, we identified groups of genes that were differentially expressed across mouse strains and therefore potentially causal for the variation in HSPC traits.

MicroRNAs are evolutionary conserved small (~22-nucleotide) non-coding RNAs that fine-tune gene expression by base-pairing with target mRNAs, leading to mRNA destabilization or translational repression.¹⁰⁻¹² Each microRNA can coordinately target hundreds of different mRNAs,¹³ and each mRNA can harbor multiple microRNA target sites, creating complex microRNA/mRNA regulatory circuitries. A growing body of evidence has implicated specific microRNAs in the regulation of HSPC fate. For example, miR-181, miR-150, and the miR-17~92 cluster have proven to be essential for lymphoid development,¹⁴⁻¹⁸ miR-223 for myeloid development,^{19,20} and miR-155 for both lymphoid and myeloid development.²¹⁻²³ Dysregulated expression of these and other microRNAs has been shown to contribute to the pathogenesis and progression of hematologic malignancies.^{24,25}

To determine whether the variation in gene expression and HSPC traits across BXD mouse strains could be due to natural variation in microRNA expression, we embarked on a genome-wide microRNA expression study of Lin⁻Sca-1⁺c-Kit⁺ (LSK), Lin⁻Sca-1⁻c-Kit⁺ (LS⁻K⁺), erythroid and myeloid cells isolated from the B6 and D2 parental mouse strains. We identified both cell type-dependent and mouse strain-dependent microRNAs. Interestingly, we discovered an evolutionary conserved cluster of microRNAs (99b/let-7e/125a) that is most highly expressed in HSPCs and that is differentially expressed between mouse strains. To assess whether the differential expression of this cluster could be functional we overexpressed it in HSPCs, and found that this conferred a competitive advantage to these cells. However, we also found that mice reconstituted with these cells developed myeloproliferative neoplasms (MPNs). Finally, we identified

the candidate functional downstream targets through which the miR-cluster may be able to modulate HSPC fate.

METHODS

Mice

Female B6 (CD45.2) and D2 mice were purchased from Harlan and housed under clean conventional conditions. Female and male B6.SJL (CD45.1) mice were bred at the Central Animal Facility of the University of Groningen. All animal experiments were approved by the Groningen University Animal Care Committee.

6

Cell purification

Bone marrow (BM) cells were flushed from the femurs and tibias of three to five mice and pooled. After standard erythrocyte lysis, nucleated cells were stained with a panel of Alexa Fluor 700-labeled lineage-specific antibodies (containing antibodies to CD3 (clone 17A2), CD11b (Mac1; clone M1/70), CD45R/B220 (clone RA3-6B2), Gr-1 (Ly-6G and Ly-6C; clone RB6-8C5) and TER-119 (clone TER-119)), fluorescein isothiocyanate (FITC)-labeled anti-Sca-1 (clone E13-161.7), phycoerythrin (PE)-labeled anti-c-Kit (clone 2B8), phycoerythrin-cyanine-7 (PE-Cy7)-labeled anti-TER-119 (clone TER-119), and allophycocyanin (APC)-labeled anti-Gr-1 (clone RB6-8C5). Antibodies were purchased from BioLegend. Cells were purified using a MoFlo flowcytometer (BeckmanCoulter) and were immediately collected in RNeasy Protect cell reagent (Qiagen) for microRNA expression studies or RNA lysis buffer (Qiagen) for gene expression studies. Triplicates were generated for each of the 8 conditions (4 cell types, 2 mouse strains).

MicroRNA expression analysis

Total RNA was isolated using the miRNeasy Mini Kit (Qiagen) in accordance with the manufacturer's protocol. 100 ng of total RNA was labeled and hybridized to Agilent 8x15K Mouse miRNA arrays, based on version 11 release of Sanger Inst miRBASE. Labeling, hybridization and washing were performed by Oxford Gene Technology using the Agilent Mouse microRNA Microarray Kit. Data were extracted using Agilent Feature Extraction Software version 9.5.3.1. Data were thresholded at 1, log₂-transformed and normalized to the 75th percentile (as recommend by Agilent) using GeneSpring GX11.0 (Agilent). On the basis of Principle Component Analysis, 1 sample was removed from further analysis. The starting data set represented 577 non-control probes. False positives were excluded by only pursuing with those probes that were flagged as present in at least 2 out of 3 replicates in any 1 out of 8 conditions. The quality-filtered expression values were median-centered per microRNA and clustered (Euclidean distance, complete linkage) using Genesis.²⁶ Signatures were identified using

a manually chosen threshold. Two-way analysis of variance (ANOVA) was applied to identify cell type-dependent and mouse-strain-dependent microRNAs. A Benjamini-Hochberg false discovery rate correction was applied to control for multiple testing.

Retroviral vectors

Genomic DNA was isolated from B6 cells and a 1.2 Kb region spanning the miR-99b, let-7e and miR-125a stem loop sequences was amplified (forward primer: 5'-CCTCGAGTGGACTGAGGAGAATTGAGTGCAAG-3', reverse primer: 5'-GCAATTGTGCCCTGAAGATCAGCAGGAAC-3'; Biolegio). The resulting PCR product was extracted from gel, TOPO-cloned into PCR4, cut from PCR4 using *XhoI* and *MunI*, and subcloned into the *XhoI-EcoRI* site of the MXW-pPGK-IRES-EGFP vector (a Murine Stem Cell Virus-based vector).²⁷ The miR-155 gene product and the MXW-pPGK-IRES-EGFP vector were a kind gift from Prof. Chen, Stanford University School of Medicine, Stanford, CA. Both the miR-cluster and miR-155 fragments were sequence verified (StarSeq; see Supplementary Information).

Retroviral overexpression of microRNAs in primary BM cells

Primary BM cells were isolated from B6 mice 4 days post intraperitoneal injection of 150 mg/kg 5-fluorouracil (Pharmachemie Haarlem), and cultured in StemSpan (StemCell Technologies) supplemented with 10% FCS, 300 ng/mL recombinant mouse stem cell factor (rmSCF; Peprotech), 20 ng/mL rmlL11 (R&D systems), 1 ng/mL Flt3 ligand (Amgen), penicillin and streptomycin. BM cells were transduced by transfecting Phoenix ecotropic packaging cells with 1-2 μ g of pDNA (MXW empty vector, MXW-miR-cluster 99b/let-7e/125a and MXW-miR-155) and 3-6 μ l Fugene HD (Roche). Virus-containing supernatant harvested 48 and 72 hours later was used to transduce $4-6 \times 10^5$ BM cells per 3.5 cm well. Three independent transductions were performed per condition per experiment. Five days after the first transduction, viable (negative for propidium iodide) EGFP⁺ cells were collected and tested in *in vitro* assays and collected in RNA lysis buffer (Qiagen) for gene expression studies. Non-sorted cells were tested in an *in vivo* BM transplantation setting.

Quantitative PCR validation

We converted selected microRNAs to cDNA using the TaqMan MicroRNA Reverse Transcription Kit (Applied Biosystems). Next, real-time PCR with Taqman MicroRNA assays (Applied Biosystems) was performed using the iCycler system (Bio-Rad). The following assays were used: hsa-miR-99b (assay ID 000436), hsa-let-7e (002406), hsa-miR-125a-5p (002198), mmu-miR-155 (002571), and snoRNA202 (001232). The comparative delta-delta C_t approximation method was used to analyze relative changes in gene expression.²⁸ Each of the three

samples was analyzed in triplicate and normalized to the endogenous control snoRNA202.

CFU-GM and CAFC assays

A granulocyte-macrophage colony-forming unit (CFU-GM) assay was performed to determine the number of progenitors in the transduced BM population. EGFP⁺ cells were plated in methylcellulose medium supplemented with 100 ng/ml recombinant mouse stem cell factor (rmSCF; Peprotech) and 20 ng/ml recombinant murine granulocyte-macrophage colony-stimulating factor (rmGM-CSF; R&D systems). After 6 days CFU-GM colonies were scored. Replating was performed to determine progenitor self-renewal capacity. The colonies from one plate were collected, washed three times with PBS, and all the cells were plated in new methylcellulose for an additional week. The cobblestone-area forming cell (CAFC) assay was performed as previously described.²⁹ In this assay, early-appearing cobblestones (day 7) are considered to represent progenitors and late-appearing cobblestones (day 35) are considered to represent stem cells.³⁰ CAFC frequencies were calculated using L-Calc (StemCell Technologies). For both of the *in vitro* assays at least two independent experiments were performed.

Primary BM transplantation

B6 BM cells were transplanted into lethally irradiated (9,5Gy, IBL 637 ¹³⁷Cs γ -source, CIS Biointernational, Gif-sur-Yvette, France) B6.SJL recipients without prior sorting for EGFP expression. Per condition 9 mice (3 independent transductions x 3 mice) were transplanted with 5×10^6 cells each. At several time points post-transplantation blood was drawn from the retro-orbital plexus and blood cell numbers were counted using a Medonic hematology analyzer (Boule Medical). Subsequently, erythrocytes were lysed in ammonium chloride solution and the remaining cells were stained with PE-labeled anti-CD45.2 (B6; clone 104), Pacific blue-labeled anti-CD45.1 (B6.SJL; clone A20), APC-Cy7-labeled anti-Gr-1 (clone RB6-8C5), PE-Cy7-labeled anti-CD11b (Mac1; clone M1/70), Pacific orange-labeled anti-CD45R/B220 (RA3-6B2; Invitrogen, Caltag Laboratories), and APC-labeled anti-CD3 ϵ (clone 145-2C11). Antibodies were purchased from BioLegend, unless otherwise specified. Data were acquired using an LSR II (BD Biosciences) and analyzed using FlowJo software (Tree Star). Three of the miR-cluster mice were sacrificed prior to the onset of morbidity (no signs of physical illness yet) at ~20 weeks post-transplantation, together with 3 empty vector and 3 miR-155 mice. Four miR-cluster mice were sacrificed when they were moribund (6, 16, 19, and 21 weeks post-transplantation), 2 of which together with an empty vector mouse. Two miR-cluster mice were found dead at 15 and 30 weeks, preventing determination of the cause of death. Peripheral blood was drawn from the retro-orbital plexus. Bone marrow cells were isolated from the 2 femurs, 2 tibias, and pelvic bones (and in some cases the spine and sternum). Single-cell suspensions

were obtained by crushing the bones and filtering the resulting cell suspension. Spleens, livers, and lungs were isolated, and made into single cell suspensions. Cell numbers were counted using a Medonic hematology analyzer. Cytospin preparations were made from BM, spleen, liver and lung cells, after which May-Grünwald-Giemsa staining was performed. Also, these cells were stained using antibodies against CD45.2, CD45.1, Gr-1, CD11b, CD45R/B220, and CD3ε (as described in the previous section) and analyzed on an LSR II. Of each of the 3 non-morbid miR-cluster mice and their empty vector and miR-155 controls viable EGFP⁺ cells were tested in CAFC assays, after which the miR-cluster and empty vector cells were also tested in a secondary BM transplantation.

Secondary BM transplantation

EGFP⁺ CD45.2⁺ BM cells from the primary miR-cluster and empty vector mice were transplanted in various cell doses in competition with freshly isolated wild type CD45.1⁺ BM cells in lethally irradiated recipients (CD42.2). Transplanted ratios were 4:1 (test to freshly isolated) for the empty vector-transduced cells and 4:1 and 1:2 for the miR-cluster-transduced cells. For each of these ratios 15 mice (3 individual mice x 5 mice) were transplanted with a total of 5x10⁶ cells each. The competitive repopulation index (CRI) was calculated using the following formula: $CRI = [(\%EGFP^+ CD45.2^+ \text{ cells} / \%CD45.1^+ \text{ cells}) / (\text{ratio of } EGFP^+ \text{ cells} / CD45.1^+ \text{ cells transplanted})]$. Blood samples were taken on a 4 to 6 week basis to determine donor chimerism and CRI.

Downstream target analysis

Genome-wide gene expression was assessed in two different sample types. First, BM samples transduced with empty vector, miR-cluster and miR-155 were profiled. Second, the four developmentally related hematopoietic cell types isolated from B6 and D2 mice that were profiled for microRNAs were also analyzed for genome-wide gene expression (independent samples). All samples were analyzed in triplicate. Total RNA was isolated using the RNeasy Mini Kit (Qiagen), after which RNA concentration, quality and integrity were measured using the Experion Automated Electrophoresis System (Bio-Rad). RNA was amplified using the Illumina TotalPrep RNA Amplification Kit (Ambion/Applied Biosystems) and hybridized to Sentrix MouseWG-6 v2.0 expression beadchips (Illumina) according to the manufacturer's instructions. Hybridization and washing were performed by our in-house Genome Analysis Facility. Scanning was carried out on the iScan System (Illumina). Data were quality checked and extracted using GenomeStudio software (Illumina), without normalization or background subtraction. Data were thresholded at 1, log₂-transformed and quantile normalized using GeneSpring GX11.0 (Agilent). The starting data set represented 45,281 probes. False positives were excluded by only pursuing with those probes that were flagged as marginal or present in all replicates in any 1 out of 3 or 8 conditions (depending

on experiment type; default detection p-values cut-off 0.8 for present and 0.6 for absent were used for flags). Subsequently, a joint analysis of microRNA and gene expression data was performed in GeneSpring. Predicted (conserved) downstream microRNA targets were imported from the TargetScan database.³¹

URLs

All raw data were deposited in the NCBI Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>). All processed microRNA and mRNA expression data on the four cell types from B6 and D2 mice were deposited on HemDb (<http://hemdb.org/>; currently still password-protected; username: guest, password: hemdb).

6

RESULTS

Identification of cell type-dependent microRNA signatures

To identify cell type-dependent and mouse strain-dependent microRNAs in the hematopoietic system, we embarked on a microarray-based microRNA profiling study of four developmentally related hematopoietic cell types isolated from the BM of B6 and D2 mice. We hybridized total RNA isolated from purified LSK multilineage cells, committed LS⁻K⁺ cells, erythroid TER-119⁺ cells and myeloid Gr-1⁺ cells (Figure 1A) to Agilent microRNA arrays and evaluated microRNA expression levels. In total, we analyzed 23 samples representing 4 different cell types and 2 different mouse strains. Of the 577 microRNAs profiled, 147 were expressed in at least 1 of these 8 conditions. To study the relationship between the samples, as well as the underlying patterns of microRNA expression, we applied an unsupervised two-way hierarchical clustering method using the 147 quality-filtered microRNAs. This analysis showed that the cell type effect greatly exceeded the mouse strain effect (Figure 1B, top tree), and revealed 8 distinct microRNA signatures (Figure 1B, left tree; Table S1). Of special interest is signature I consisting of microRNAs that are most highly expressed in LSKs and down-regulated upon differentiation. Signature I consists of miR-125a-5p, miR-125b-5p, miR-126-3p, miR-130a, miR-155, miR-181d, miR-196b, miR-203, miR-222, miR-31, miR-99a, miR-99b, and let-7e. Considering their expression pattern, these microRNAs may represent important factors that keep cells from leaving the stem cell state. Signature II consists of microRNAs that are most highly expressed in LS⁻K⁺ cells, signatures III and IV are specific to erythroid cells, and signatures VI and VIII to myeloid cells. Signature VII consists of microRNAs that are expressed in all cell types except for LSKs, and may therefore consist of microRNAs that are important for lineage specification, commitment and differentiation. Finally, signature V does not show a clear expression preference. We confirmed well-known cell-type-specific microRNAs: e.g. miR-196b in signature I,³² miR-451 in

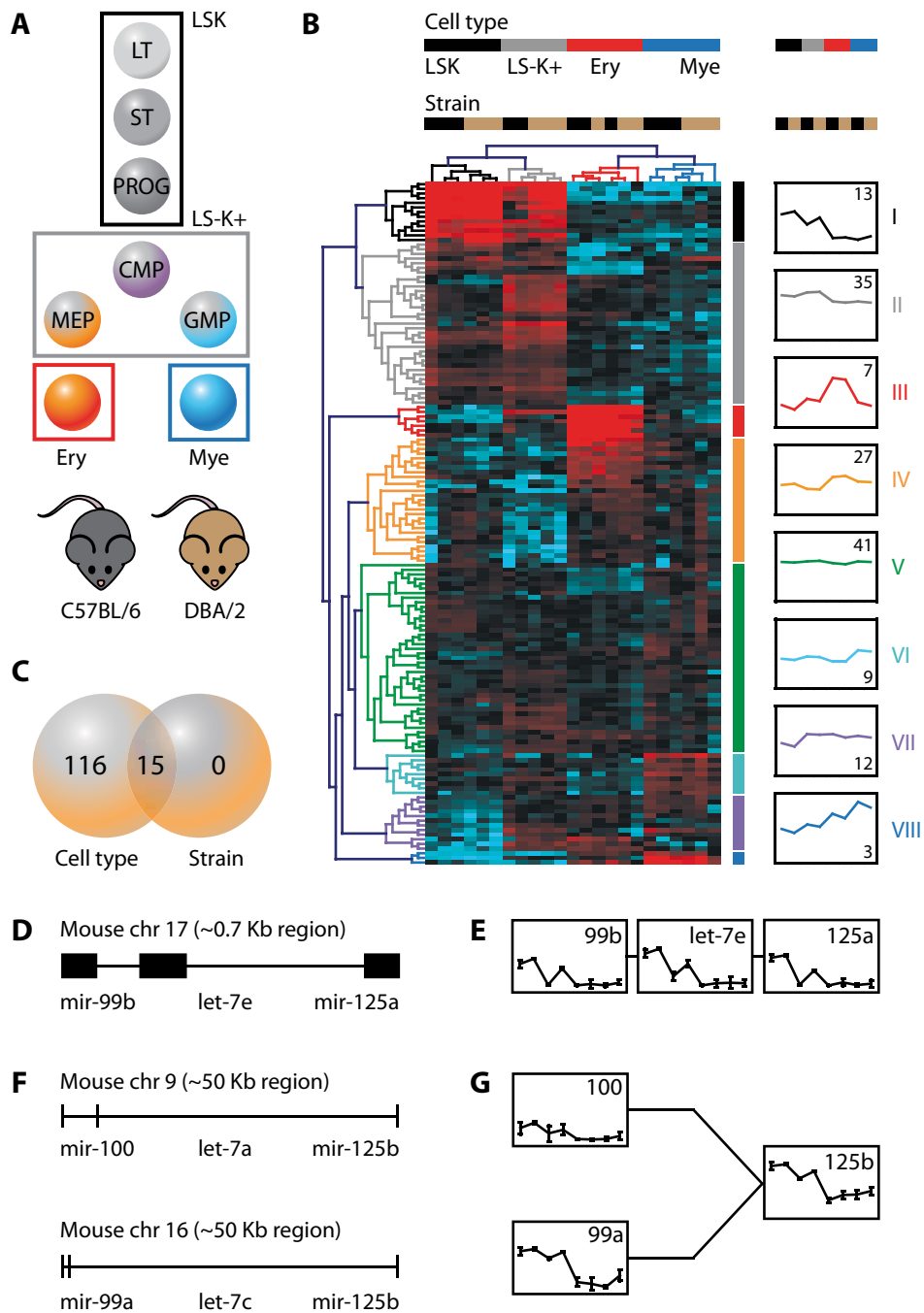
erythroid signature III,³³ and miR-223 in myeloid signature VIII.^{19;20} The entire microRNA expression dataset is available (Table S1 and in GEO), and queries for specific microRNAs can be done on www.hemdb.org.

Identification of mouse strain-dependent miR-cluster 99b/let-7e/125a

An analysis of variance (2-way ANOVA, corrected p-value <0.05) yielded 131 differentially expressed microRNAs, all of which showed a cell type effect and of which 15 also showed a mouse strain effect (Figure 1C, Table S1). In total, 4 of the 15 mouse-strain-dependent microRNAs were present in signature I, namely miR-125a, miR-125b, miR-130a and miR-99b. They were all most highly expressed in the primitive cell compartment and down-regulated upon differentiation and generally higher expressed in D2 cells compared to B6 cells. These 2 criteria qualify those microRNAs as potential candidates regulating HSPC traits. Interestingly however, 2 out of these 4, miR-125a and miR-99b are part of an evolutionary conserved microRNA cluster that is localized on chromosome 17: miR-cluster 99b/let-7e/125a (Figure 1D). The third member of this cluster, let-7e, shows an expression pattern comparable to the other 2 members (Figure 1E), but its strain effect did not reach statistical significance (corrected p-value = 0.064). In addition, miR-125b is present in two paralogous microRNA clusters that are located on chromosomes 9 and 16 (Figure 1F). Since the three paralogous microRNA clusters have identical seed sequences and therefore identical (predicted) downstream targets, it is the cumulative expression pattern of these three clusters that should be taken into consideration. The chromosome 9 cluster consists of miR-100, let-7a and miR-125b and the chromosome 16 cluster consists of miR-99a, let-7c and miR-125b. Although not all microRNA cluster members display a significant mouse strain effect, they all follow the same expression trend across cell types and mouse strains (Figure 1G). This view corroborates previous observations that proximal pairs of microRNAs (separated by <50 Kb) are generally co-expressed across tissues.³⁴

Overexpression of miR-cluster 99b/let-7e/125a retains HSPCs in a primitive state

To assess whether differential expression of these microRNA clusters could have functional consequences for HSPCs, and could be causal for the phenotypic differences between HSPCs from B6 and D2 mice, we performed a gain-of-function study for one of the three paralogous clusters: miR-cluster 99b/let-7e/125a. We cloned the entire cluster (in its natural genomic context) in a retroviral vector containing a constitutively active pol II promoter. We overexpressed the entire miR-cluster 99b/let-7e/125a and also miR-155 (known to affect HSPC characteristics and therefore used as a positive control²¹) in BM from 5-FU-treated B6 mice (Figure 2A). Transduced cells were sorted and overexpression of the individual



members of the miR-cluster and miR-155 was validated using qPCR (Figure 2B). Concurrently, these cells were used to initiate colony-forming unit- granulocyte/macrophage (CFU-GM) replating assays and cobblestone area-forming cell (CAFC) assays. The CFU-GM replating assay, in which we tested for myeloid progenitor self-renewal, revealed that both the miR-155 and the miR-cluster overexpressing cells had acquired the capacity to form secondary colonies upon replating (Figure 2C). The CAFC assay, a surrogate *in vitro* assay to quantify the number of HSPCs,^{29,30} revealed even more striking differences. At day 7, which is the time point at which normally progenitor cells show activity, no differences were observed between the control and microRNA-overexpressing cells. However, at day 35, which is the time point at which stem cells are considered to show activity, a striking difference in the number of CAFCs was observed. The miR-155 overexpressing cells gave ~15-fold more CAFCs (1 in 26.000 for miR-155 versus 1 in 450.000 for control), whereas the miR-cluster overexpressing cells yielded a striking ~6.500-fold more CAFCs (1 in 70 for miR-cluster versus 1 in 450.000 for control) (Figure 2D). For the miR-cluster, a substantial amount of CAFC activity remained even up till day 70 (at that point still 1 in 50.000, data not shown).

To summarize, BM cells overexpressing miR-cluster 99b/let-7e/125a or miR-155 showed increased replating capacity in CFU-GM assays and increased day-35 CAFC activity. Collectively, these data suggest that overexpression of the miR-cluster retains cells in a primitive state.

- ◀ **Figure 1. Genetic screen identifies cell type and mouse strain-dependent microRNAs.** (A) MicroRNA expression was evaluated in four developmentally related cell types isolated from the BM of C57BL/6 and DBA/2 mouse strains. (B) Hierarchical clustering was performed using the 147 quality-filtered probes (Euclidean distance, complete linkage). Samples are in columns, microRNAs in rows. For each probe, data were median-centered, with the lowest and highest intensity values in blue and red, respectively. Of each of the 8 microRNA signatures the average expression across cell types and mouse strains is shown (y-axis represents log₂ expression ranging from -4 till 8). The number in the graphs indicates the number of microRNAs per signature. (C) Venn diagram showing the number of differentially expressed microRNAs between cell types and mouse strains. (D) An evolutionary conserved microRNA cluster, consisting of miR-99b, let-7e and miR-125a, located on mouse chromosome 17. (E) Expression of miR-cluster 99b/let-7e/125a members across cell types and mouse strains (axes: as in B). Shown is the mean +/- standard deviation. (F) Paralogous microRNA clusters on chromosomes 9 and 16. (G) Expression of miR-100, miR-99a and miR-125b across cell types and mouse strains (axes: as in B). For miR-125b only the cumulative expression of the chromosome 9 and 16 cluster could be assessed. Expression of let-7a and let-7c is not shown, as only the cumulative expression from multiple different genomic locations (including other chromosomes than 9, 16 and 17) could be assessed. Shown is the mean +/- standard deviation. LT, long-term; ST, short-term; CMP, common myeloid progenitor; MEP, megakaryocyte-erythrocyte progenitor; GMP, granulocyte-macrophage progenitor.

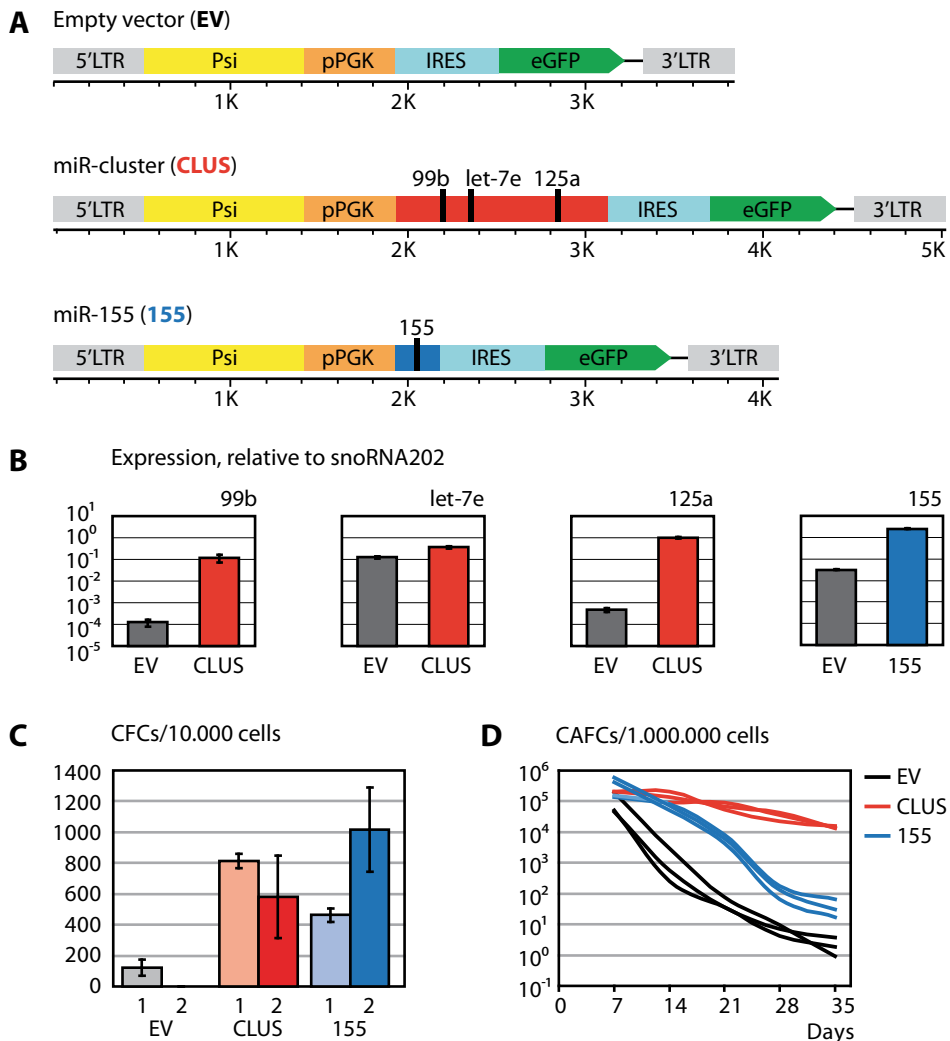


Figure 2. Overexpression of miR-cluster 99b/let-7e/125a fixes HSPCs in a primitive state. (A) Schematic representation of the retroviral vectors used to overexpress miR-cluster 99b/let-7e/125a and miR-155. (B) Quantitative RT-PCR data showing the expression levels of the miR-cluster members and miR-155 upon overexpression relative to the endogenous control snoRNA202. Shown is the mean \pm standard deviation. (C) Colony-forming unit-granulocyte/macrophage data showing primary (1) and secondary (2) colony numbers. Shown is the mean \pm standard deviation. (D) Cobblestone area-forming cell data showing the number of HSPCs.

Sustained expression of miR-cluster 99b/let-7e/125a in HSPCs results in MPNs

We next assayed the microRNA-overexpressing cells in a long-term competitive repopulation experiment *in vivo*. To achieve this, miR-cluster 99b/let-7e/125a and miR-155 were overexpressed in B6 cells (CD45.2) after which 5×10^6 non-sorted cells were transplanted into lethally irradiated B6.SJL recipients (CD45.1). At 10 weeks post-transplantation, blood cell counts revealed a significant reduction in white blood cells, red blood cells and platelets for mice reconstituted with miR-cluster-overexpressing cells. Although the decrease was less dramatic, a similar effect could be observed for the mice reconstituted with miR-155-overexpressing cells (Figure 3A). At this point in time, chimerism was assessed by quantifying the percentage of EGFP⁺ cells within the CD45.2⁺ donor fraction. Whereas within 10 weeks the chimerism levels of the control mice remained relatively stable at 50%, those of the miR-155 mice increased from 15% to 45%, and those of the

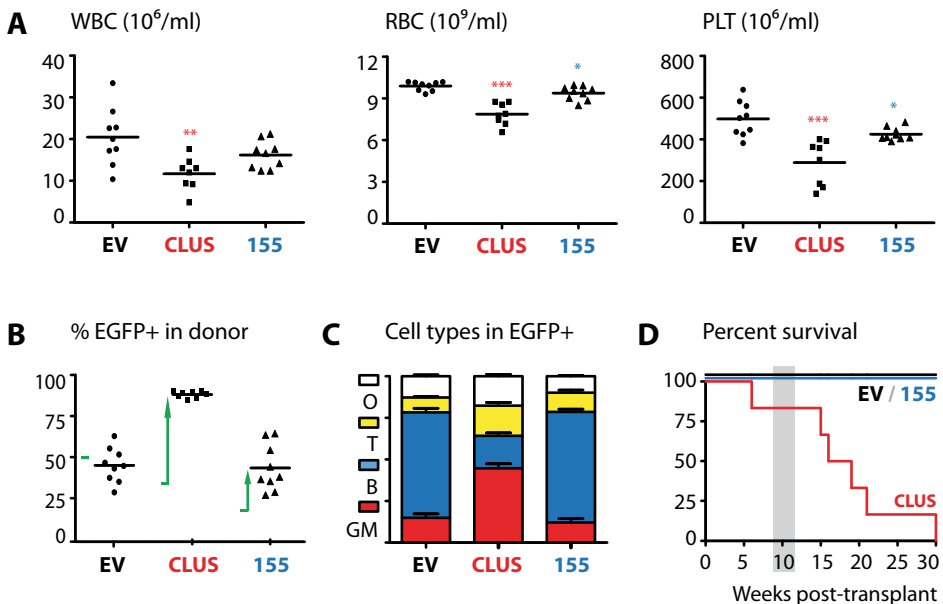
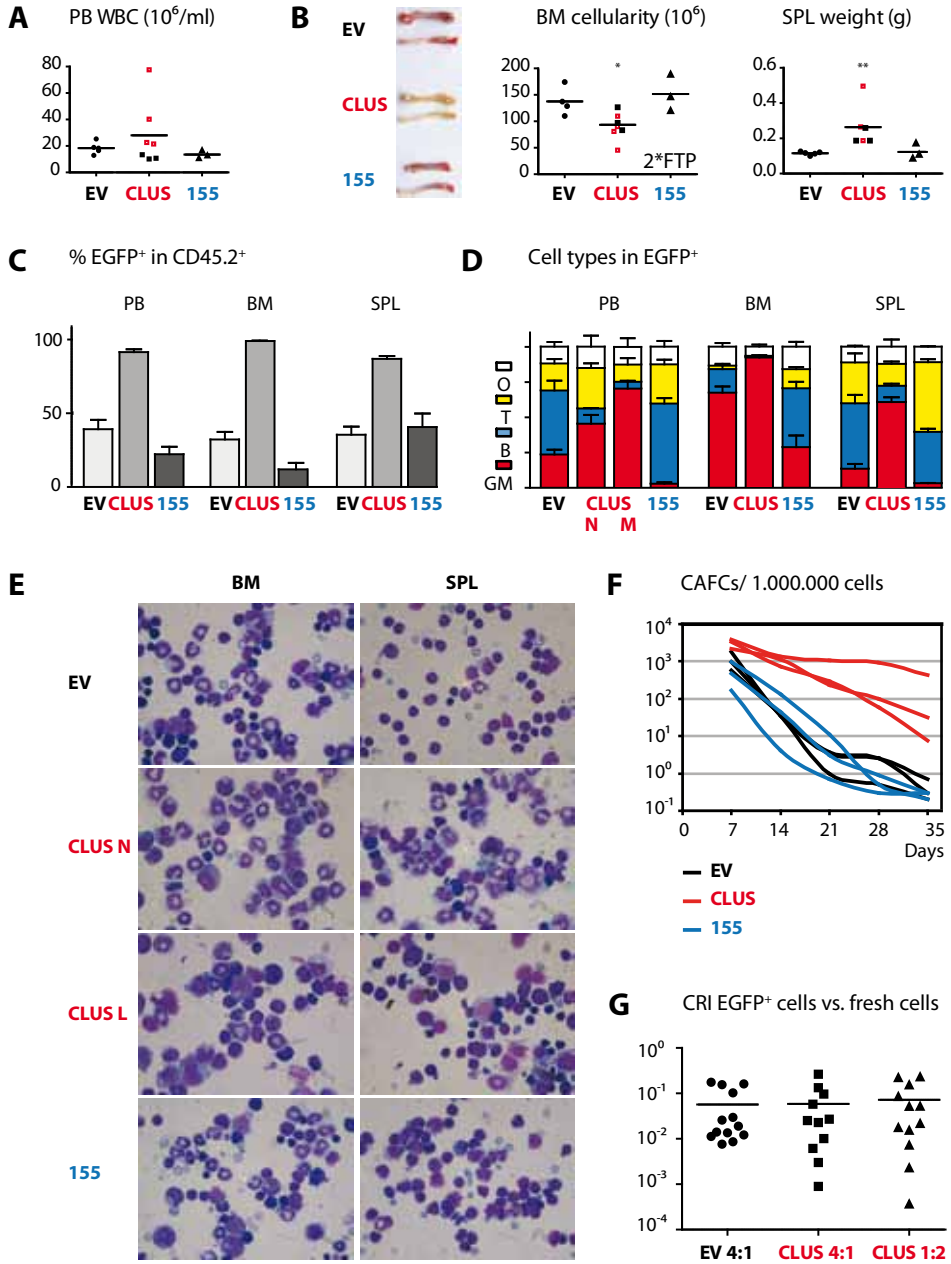


Figure 3. Sustained expression of miR-cluster 99b/let-7e/125a disturbs hematopoiesis.

(A) White blood cell, red blood cell and platelet counts at 10 weeks post-transplantation ($n = 8/9$ mice per group). P values (**, $0.001 < p < 0.01$; ***, $p < 0.001$) are displayed (Mann-Whitney test). (B) Chimerism levels were assessed at 10 weeks post-transplantation by analyzing the percentage of EGFP⁺ cells in the donor fraction ($n = 8/9$ mice per group). The arrows indicate the increase in chimerism levels over the first 10 weeks post-transplantation. (C) Cell type distribution in the blood as assessed by FACS ($n = 8/9$ mice per group). Shown is the mean \pm SEM. (D) Survival of mice ($n = 6$ mice per group). The grey box indicates the time point analyzed in the other panels. WBC, white blood cells; RBC, red blood cells; PLT, platelets; GM, granulocytes/macrophages; B, B-lymphocytes; T, T-lymphocytes; O, other cell types.



miR-cluster mice increased from 35% to 90%, indicating that in the latter two cases the microRNA-transduced cells had a competitive advantage over the non-transduced cells (Figure 3B). Detailed FACS analysis revealed that the EGFP⁺ cells in the miR-cluster mice were enriched for granulocytes/macrophages (defined as being Gr-1⁺ and/or Mac-1⁺) at the expense of the number of B lymphocytes (B220⁺) (Figure 3C). Clearly, hematopoiesis was disturbed in these mice, and this appeared to be not without consequences, as all miR-cluster mice died within 30 weeks post-transplantation (Figure 3D).

To obtain more insight into the type of hematologic malignancy in the miR-cluster mice, we performed pathological analyses on 3 mice prior to the onset of morbidity at 20 weeks post-transplantation and on 4 mice when they were moribund (6, 16, 19, and 21 weeks post-transplantation). We compared our findings with analyses on mice reconstituted with empty vector and miR-155-transduced cells. The peripheral blood analysis revealed that all the moribund mice displayed higher WBC counts compared to the non-morbid mice. Among these moribund mice, there was one clearly leukemic (and one potentially pre-leukemic) mouse (Figure 4A). Compared to the empty vector mice, the miR-cluster mice had whitish bones, lower BM cellularities, and increased spleen sizes

◀ **Figure 4. Sustained expression of miR-cluster 99b/let-7e/125a induces MPNs with occasional progression to leukemia.** (A) PB WBC counts for empty vector, miR-cluster and miR-155 mice at time of sacrifice. Closed/black squares represent non-morbid miR-cluster mice, whereas open/red squares represent moribund miR-cluster mice. (B) Photographs of femur and tibia bones, BM cellularity (representing 2 femurs, 2 tibias and pelvic bones) and spleen weight for empty vector, miR-cluster and miR-155 mice at time of sacrifice. Closed/black squares represent non-morbid miR-cluster mice, whereas open/red squares represent moribund miR-cluster mice. P values (*, 0.01 < p < 0.05; **, 0.001 < p < 0.01) are displayed (Mann-Whitney test). (C) Percentage EGFP⁺ cells in PB, BM and spleen (n=5 for EV, n=6 for CLUS, n=3 for 155). Non-morbid and moribund miR-cluster mice showed similar percentages and were therefore combined. Shown is the mean ± SEM. (D) Cell type distribution in PB, BM and spleen of empty vector, miR-cluster and miR-155 mice as assessed by FACS. Shown is the mean ± SEM. Data for non-morbid and moribund mice are shown separately for PB (n=5 for EV, n=3 for non-morbid (N) CLUS, n=3 for moribund (M) CLUS, n=3 for 155) and combined for BM and spleen (n=5 for EV, n=6 CLUS, n=3 for 155), because the cell type distributions between these mice differed in PB, but not in BM and spleen. (E) Representative May-Grünwald-Giemsa-stained cytospin preparations from BM and spleen cells. For the miR-cluster, two pictures are shown: one representative of a non-morbid mouse (N) and one of the clearly leukemic mouse (L). (F) Cobblestone area-forming cell (CAFC) data showing the number of HSPCs in the EGFP⁺ BM fraction of 3 individual empty vector, (non-morbid) miR-cluster and miR-155 mice at ~20 weeks post-transplantation. (G) Competitive repopulation indexes (CRI) calculated for EGFP⁺ BM cells isolated from 3 individual empty vector and (non-morbid) miR-cluster mice at ~20 weeks post-transplantation, compared to freshly isolated BM cells. Transplanted ratios were 4:1 (test to freshly isolated) for the empty vector-transduced cells and 4:1 and 1:2 for the miR-cluster-transduced cells. Calculations were done 16 weeks after secondary transplantation. For the miR-cluster-transduced cells 1 outlier with a CRI of 1.3 was removed from this analysis. WBC, white blood cells; GM, granulocytes/macrophages; B, B-lymphocytes; T, T-lymphocytes; O, other cell types.

(Figure 4B). Subsequent FACS analyses revealed that the PB, BM and spleens of the miR-cluster mice were dominated by EGFP⁺ cells (Figure 4C) that were mostly of myeloid (defined as being Gr-1⁺ and/or Mac-1⁺) origin (Figure 4D). This latter finding was corroborated by morphological analyses of MGG-stained cytospin preparations (Figure 4E). The livers and lungs of the miR-cluster mice were also infiltrated by these EGFP⁺ myeloid cells (data not shown). Taken together, these data show that all miR-cluster mice suffered from MPNs that in some cases progressed to leukemia. In the miR-155 mice an opposite phenotype could be observed, as the EGFP⁺ cells in the PB, BM and spleens of these mice were mostly of lymphoid (defined as being B220⁺ or CD3⁺) origin (Figure 4D and 4E).

Competitive advantage provided by miR-cluster 99b/let-7e/125a decreases over time

To determine the quantity of HSPCs in the BM compartments of the non-morbid miR-cluster mice and their empty vector and miR-155 controls, we performed CAFC assays on EGFP⁺ BM cells. To recapitulate: before we transplanted these cells we found ~15-fold more day-35 CAFCs for miR-155-transduced cells and a striking ~6,500-fold more day-35 CAFCs for the miR-cluster-transduced cells. At 20 weeks post-transplantation the miR-cluster-transduced cells still exhibited more day-35 CAFC activity compared to empty vector-transduced cells, whereas this was not the case for the miR-155-transduced cells (Figure 4F). However, it should be noted that for the miR-cluster-transduced cells the fold increase in day-35 CAFC activity dropped from ~6,500 to ~400 in this 20 week time frame *in vivo*.

To determine the quality of the HSPCs in the BM compartments of the miR-cluster and empty vector mice, we performed a secondary BM transplantation. We evaluated the *in vivo* repopulating ability of CD45.2⁺ EGFP⁺ cells isolated from the BM of the primary recipients by transplanting them in competition with freshly isolated CD45.1⁺ BM cells. We next determined the competitive repopulation index (CRI), a relative measurement of the competitive ability of the CD45.2⁺ EGFP⁺ test cells to that of the freshly isolated CD45.1⁺ BM cells. A CRI value of 1 would by definition mean equal repopulation potential of both cell populations. At 16 weeks post-transplantation, we found CRI values of ~0.06 for both the empty vector and miR-cluster-transduced test cells (i.e. their repopulation potential was ~17-fold lower compared to freshly isolated cells) (Figure 4G). A CRI value of 0.06 is around what would be expected for control cells after a single serial transplantation.³⁵ These data indicate that whereas the miR-cluster-transduced cells may have had a profound initial competitive advantage in the primary recipients, this was lost in the secondary recipients.

To summarize, we found that overexpressing the miR-cluster in HSPCs indeed conferred a competitive advantage to engrafting hematopoietic cells in an *in vivo* transplant setting, ultimately resulting in MPNs with occasional progression to leukemia. Yet, this competitive advantage appeared to be of temporary nature.

In search of functional miR-cluster 99b/let-7e/125a targets

To understand the molecular mechanisms by which miR-cluster 99b/let-7e/125a is able to module HSPC fate, it is essential to identify the genes regulated by this microRNA triplet. Many different computational algorithms have been developed which can be used for microRNA target prediction. TargetScan is one of the most widely used algorithms and predicts 48, 712, and 572 evolutionary conserved targets for miR-99b, let-7e, and miR-125a, respectively.³¹ However, it remains challenging to select the functionally relevant targets from the long list of predicted ones.

Our first strategy to narrow down the list of predicted targets to functionally relevant ones was to select only those targets whose expression reacts on changing microRNA expression levels. To achieve this for the miR-cluster and miR-155 (included as a control again), we performed Illumina WG6 gene expression arrays on empty vector, miR-cluster and miR-155-transduced cells. Viable EGFP⁺ cells were FACS-purified 5 days after the first transduction (the exact time point at which the before-mentioned *in vitro* and *in vivo* assays were initiated). Of the 45,281 probes, 25,808 were expressed in at least 1 of the 3 conditions (empty vector, miR-cluster, and miR-155). In this list of quality-filtered probes we identified 912 predicted targets of the miR-cluster and 164 of miR-155. We then visualized the expression levels of these predicted targets upon overexpression of the miR-cluster and miR-155 (Figure 5A). This analysis showed that the expression of most of the predicted miR-cluster targets was indeed mildly affected in the miR-cluster-transduced cells, but not in the miR-155-transduced cells. In contrast, most of the miR-155 targets were only affected in the miR-155-transduced cells, but not in the miR-cluster-transduced cells. These results confirm previous reports that the primary function of microRNAs is to fine-tune gene expression (i.e. they do not function as master regulators). From the list of predicted miR-cluster targets we subsequently selected those probes that were at least 1.2 fold differentially expressed upon overexpression of the miR-cluster. This resulted in 9 probes that were up-regulated and 71 probes that were down-regulated upon miR-cluster overexpression. These latter 71 represented 64 genes that would qualify as candidate functional targets as they represent predicted targets that are indeed down-regulated upon miR-cluster overexpression. The entire gene expression dataset is available (Table S2 and in GEO).

Our second strategy to narrow down the list of predicted targets to functionally relevant ones was to correlate the expression levels of the predicted miR-cluster targets with the expression of the miR-cluster itself in primary, unperturbed, hematopoietic cells. We reasoned that the naturally occurring variation in microRNA expression should at least have consequences on the mRNA level in order to affect any HSPC traits. Therefore, we also performed Illumina WG6 gene expression arrays on the same four developmentally related cell types isolated from B6 and D2 mouse strains that were also profiled for microRNAs. Of the

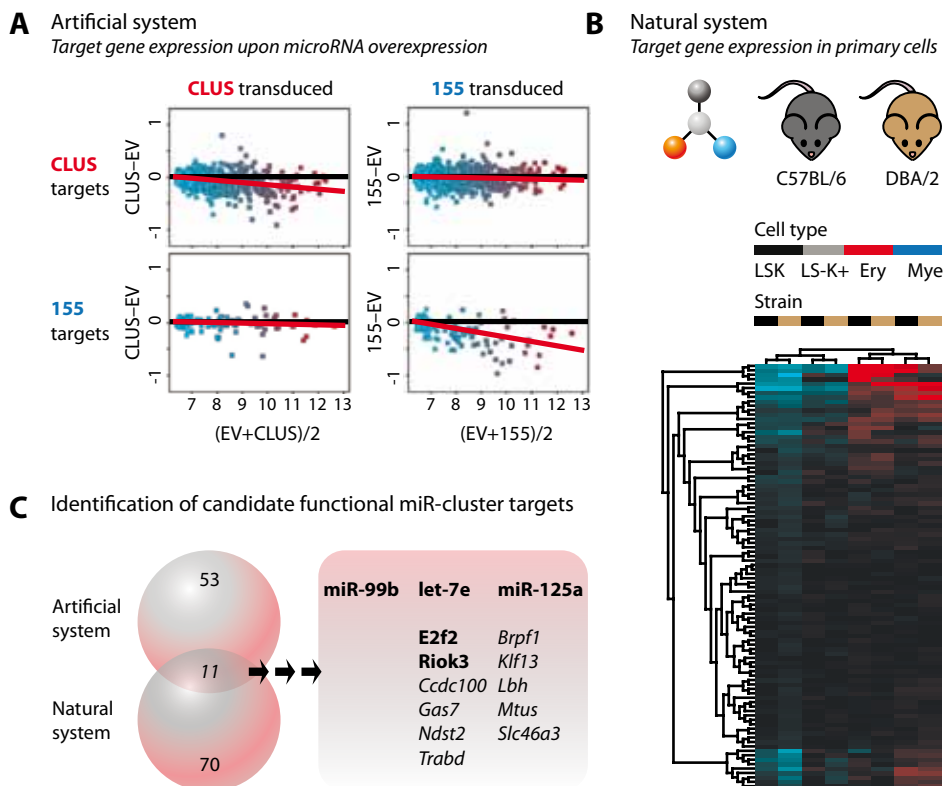


Figure 5. Identification of candidate functional miR-cluster targets. (A) Visualization of expression of all quality-filtered predicted miR-cluster and miR-155 targets in miR-cluster-transduced and miR-155-transduced cells. Shown are MvA plots comparing two samples. The log₂ ratio of each probe (expression difference) is plotted versus the log₂ mean for each probe. The red lines represent the best fit to the data. In these plots the predicted targets that are shared between the miR-cluster and miR-155 are not displayed. (B) Hierarchical clustering (Euclidean distance, complete linkage) was performed using the quality-filtered predicted miR-cluster targets that were most anti-correlated with miR-cluster expression (Pearson similarity values between -0.6 and -1.0). Conditions are in columns, mRNA probes in rows. For each probe, data were median-centered, with the lowest and highest intensity values in blue and red, respectively. (C) Venn diagram identifying 11 genes that were both down-regulated upon miR-cluster overexpression in an artificial system and anti-correlated with miR-cluster expression in the natural system. In addition, the individual members of the miR-cluster are shown with their candidate functional targets.

45,281 probes, 26,594 were expressed in at least 1 of the 8 conditions (four cell types, two mouse strains). In this list of quality-filtered probes we identified 909 predicted targets of the miR-cluster. We correlated their expression levels with the expression of the miR-cluster (average of three individual members) across the 8 different conditions, and selected the 96 most anti-correlated probes (Pearson similarity values between -0.6 and -1.0). These 96 probes represented

81 different genes. The entire gene expression dataset is available (Table S3 and in GEO). To visualize gene expression in the four cell types and two mouse strains queries for specific genes can be done on www.hemdb.org.

Finally, we combined our list of 64 genes that were 1.2 fold down-regulated upon miR-cluster overexpression with our list of 81 most anti-correlated genes in unperturbed primary cells, and identified 11 overlapping ones (Figure 5C), namely *Brpf1*, *Ccdc100*, *E2f2*, *Gas7*, *Klf13*, *Lbh*, *Mtus1*, *Ndst2*, *Riok3*, *Slc46a3*, and *Trabd*. Two genes in the list of 11 candidate functional miR-cluster targets were identified by 2 distinct probes, making them especially strong candidate targets. The first one, *E2f2*, is a well-known critical regulator of the cell cycle and has already been experimentally validated as a direct target of the let-7 family.³⁶ It has previously been shown that the combined loss of both *E2f1* and *E2f2* in mice results in severely impaired hematopoiesis with impeded S-phase progression in hematopoietic progenitors, impeded erythroid maturation, and increased apoptosis during B-cell maturation.^{37;38} Clearly, these findings are in line with at least part of our observations regarding the miR-cluster gain-of-function phenotype. Interestingly, *E2f2* is a predicted target of both let-7 and miR-155. Upon miR-155 overexpression *E2f2* was also clearly down-regulated, potentially explaining some of the similarities in the miR-cluster and miR-155 gain-of-function phenotypes (e.g. both providing HSPCs with an initial competitive advantage *in vivo*). The second candidate functional miR-cluster target that was identified by two distinct probes is *Riok3*, a relatively unknown gene predicted to be targeted by the let-7 family. Recently, it has been shown to be essential for erythroid chromatin condensation and enucleation,³⁹ suggesting that the down-regulation of this gene in our miR-cluster mice could have contributed to the observed impeded erythroid maturation. Future studies are needed to experimentally validate the candidate functional miR-cluster targets.

DISCUSSION

In this report, we set out to determine whether the observed natural variation in HSPC traits across the genetically distinct mouse strains B6 and D2 could be due to natural variation in microRNA expression. We performed a genome-wide microRNA expression study of LSK, LS⁻K⁺, erythroid and myeloid cells isolated from B6 and D2 mouse strains, and detected natural variation in microRNA expression between these two mouse strains. Of special interest was an evolutionary conserved miR-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 upon differentiation and generally higher expressed in D2 cells compared to B6 cells. We found that overexpression of this miR-cluster conferred a competitive advantage to HSPCs, but also that mice reconstituted with these cells developed MPNs that occasionally progressed to leukemia. Finally, we

identified 11 candidate functional targets through which the miR-cluster could modulate HSPC fate.

While we were performing our final experiments, Guo *et al.* reported that miR-99b, let-7e and miR-125a were enriched in long-term HSCs (LSK/CD34-/Flk2-) and that BM cells overexpressing this microRNA triplet displayed enhanced reconstitution after transplantation. More specifically, they found that overexpression of miR-125a alone, but not miR-99b or let-7e alone, could amplify the HSC pool. This result was shown to be accomplished by decreasing the level of apoptosis in hematopoietic progenitors through targeting pro-apoptotic genes such as *Bak1*.⁴⁰ In our mRNA dataset on miR-cluster-overexpressing cells we could not confirm the down-regulation of *Bak1*. This discrepancy could be due to the fact that Guo *et al.* determined the down-regulation of *Bak1* on the protein level instead of mRNA level and/or that they determined the down-regulation in cell lines instead of primary cells. Concurrently, O'Connell *et al.* reported that overexpression of miR-125b caused dose-dependent MPNs that progressed to a lethal acute myeloid leukemia (AML), and also provided evidence that miR-125b promotes hematopoietic engraftment of human HSCs.⁴¹ Soon after these two initial reports on the miR-125 family in HSCs appeared, Bousquet *et al.* reported that overexpression of miR-125b could not only cause MPNs, but also B-cell acute lymphoblastic leukemia and T-cell acute lymphoblastic leukemia. Also, they provided evidence that miR-125b confers a proliferative advantage to leukemic cells.⁴² Finally, Ooi *et al.* reported that overexpression of miR-125b in HSCs enhances their function and enriches for lymphoid-balanced and lymphoid-biased HSCs. They also described that a skewing towards the lymphoid lineage could be observed in the peripheral blood of these mice, and that a small subset of them developed a lymphoproliferative disease. In search of miR-125b targets, they performed a genome-wide mRNA expression study of miR-125b-overexpressing LSKs isolated from transplanted recipients. They selected the top list of predicted targets that were down-regulated upon miR-125b overexpression and that had a link to apoptosis, and identified *Bmf* and *Klf13* as downstream targets.⁴³ Whereas in our dataset we could not confirm the down-regulation of *Bmf* upon miR-cluster overexpression, *Klf13* did pop up in our list of 11 candidate functional targets (without even pre-selecting for apoptosis-related genes). In another study, *Klf13* has already been experimentally validated as a direct target of the miR-125 family.⁴⁴

Collectively, our data and these recent reports point to an important role for the miR-125 family in hematopoiesis. It is interesting, yet confusing, to see that all the reports on the miR-125 family in HSCs describe distinct gain-of-function phenotypes (from no- to myeloid- to lymphoid malignancies). As suggested by O'Connell *et al.* differences in overexpression levels probably underlie these phenotypic differences. The overexpression levels in our experiments were similar to those of O'Connell *et al.*, probably explaining the similarity in both of our gain-of-function phenotypes. The differences that remain between the two phenotypes

are likely to be due to miR-99b and let-7e that were in our case simultaneously overexpressed with the miR-125 family member.

An interesting observation was that overexpression of the miR-cluster conferred a striking initial competitive advantage to HSPCs, but that this advantage was lost over time. This observation relates to the differences in HSPC traits between B6 and D2 mice. First, hematopoietic recovery by B6 and D2 stem cells has been studied in a competitive transplantation setting and has uncovered interesting kinetics. Whereas D2 hematopoiesis was predominant initially, it was eclipsed by B6 hematopoiesis over time.⁴⁵ Second, 2.6% of B6 stem cells has been shown to be in S-phase of the cell cycle, versus a striking 24% of D2 stem cells.³ Together, these findings indicate that the increased proliferation rate for D2 stem cells may have provided them with an initial competitive advantage, but that this increased cycling rate also made them exhaust more rapidly. Because of the resemblance with the miR-cluster gain-of-function phenotype, we consider it plausible that the differential expression of miR-cluster 99b/let-7e/125a (and thus its targets) may at least be in part responsible for the observed variation in stem cell cycling and repopulation kinetics between B6 and D2 mouse strains.

To conclude, we uncovered the existence of natural variation in microRNA expression between mouse strains, and showed that this may contribute to the observed natural variation in HSPC traits. We anticipate that future genetic studies will shed new light on how microRNAs modulate HSPC fate and on how they themselves are regulated.

ACKNOWLEDGMENTS

We thank Mathilde Broekhuis, Jaring Schreuder, Albertina Ausema, and Bahram Sanjabi for technical assistance; Henk Moes, Geert Mesander and Roelof Jan van der Lei for assistance in cell sorting; and Brad Dykstra, Ronald van Os, Joost Kluiver, Anke van den Berg and Gerwin Huls for scientific advice. This work was supported by the Dutch Cancer Society (RUG2007-3729); the Netherlands Genomics Initiative (Horizon, 050-71-055); the Netherlands Organization for Scientific Research (VICI, 918-76-601 to G.d.H.); and by the European Community (EuroSystem, 200720).

SUPPORTING INFORMATION AVAILABLE ON REQUEST

Supplementary methods

MiR-cluster 99b/let-7e/125a insert in *Xho-EcoRI* site (sequence verified)

MiR-155 insert in *Xho-EcoRI* site (sequence verified)

Table S1. MicroRNA expression dataset on four cell types isolated from B6 and D2 mice

Table S2. Gene expression dataset on microRNA-overexpressing BM cells

Table S3. Gene expression dataset on four cell types isolated from B6 and D2 mice

REFERENCES

- De Haan G, Van Zant G. Intrinsic and extrinsic control of hemopoietic stem cell numbers: mapping of a stem cell gene. *J.Exp.Med.* 1997;186(4):529-536.
- De Haan G, Nijhof W, Van Zant G. Mouse strain-dependent changes in frequency and proliferation of hematopoietic stem cells during aging: correlation between lifespan and cycling activity. *Blood* 1997;89(5):1543-1550.
- Van Zant G, Eldridge PW, Behringer RR, Dewey MJ. Genetic control of hematopoietic kinetics revealed by analyses of allophenic mice and stem cell suicide. *Cell* 1983;35(3 Pt 2):639-645.
- Muller-Sieburg CE, Riblet R. Genetic control of the frequency of hematopoietic stem cells in mice: mapping of a candidate locus to chromosome 1. *J.Exp.Med.* 1996;183(3):1141-1150.
- Morse HC, III, Chused TM, Hartley JW et al. Expression of xenotropic murine leukemia viruses as cell-surface gp70 in genetic crosses between strains DBA/2 and C57BL/6. *J.Exp.Med.* 1979;149(5):1183-1196.
- Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC.Genet.* 2004;57.
- Gerrits A, Dykstra B, Otten M, Bystrykh L, De Haan G. Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny. *Immunogenetics* 2008;60(8):411-422.
- Bystrykh L, Weersing E, Dontje B et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat.Genet.* 2005;37(3):225-232.
- Gerrits A, Li Y, Tesson BM et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS.Genet.* 2009;5(10):e1000692.
- Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;116(2):281-297.
- Ambros V. The functions of animal microRNAs. *Nature* 2004;431(7006):350-355.
- Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 2010;466(7308):835-840.
- Lim LP, Lau NC, Garrett-Engele P et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 2005;433(7027):769-773.
- Chen CZ, Li L, Lodish HF, Bartel DP. MicroRNAs modulate hematopoietic lineage differentiation. *Science* 2004;303(5654):83-86.
- Xiao C, Calado DP, Galler G et al. MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb. *Cell* 2007;131(1):146-159.
- He L, Thomson JM, Hemann MT et al. A microRNA polycistron as a potential human oncogene. *Nature* 2005;435(7043):828-833.
- Xiao C, Srinivasan L, Calado DP et al. Lymphoproliferative disease and autoimmunity in mice with increased miR-17-92 expression in lymphocytes. *Nat.Immunol.* 2008;9(4):405-414.
- Zhou B, Wang S, Mayr C, Bartel DP, Lodish HF. miR-150, a microRNA expressed in mature B and T cells, blocks early B cell development when expressed prematurely. *Proc.Natl.Acad.Sci.U.S.A.* 2007;104(17):7080-7085.
- Johnnidis JB, Harris MH, Wheeler RT et al. Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. *Nature* 2008;451(7182):1125-1129.
- Fazi F, Rosa A, Fatica A et al. A microcircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis. *Cell* 2005;123(5):819-831.
- O'Connell RM, Rao DS, Chaudhuri AA et al. Sustained expression of microRNA-155 in hematopoietic stem cells causes a myeloproliferative disorder. *J.Exp.Med.* 2008;205(3):585-594.
- Thai TH, Calado DP, Casola S et al. Regulation of the germinal center response by microRNA-155. *Science* 2007;316(5824):604-608.
- Rodriguez A, Vigorito E, Clare S et al. Requirement of bic/microRNA-155 for normal immune function. *Science* 2007;316(5824):608-611.
- Esquela-Kerscher A, Slack FJ. Oncomirs - microRNAs with a role in cancer. *Nat.Rev.Cancer* 2006;6(4):259-269.
- Kluiver J, Kroesen BJ, Poppema S, Van den Berg A. The role of microRNAs in normal hematopoiesis and

- hematopoietic malignancies. *Leukemia* 2006;20(11):1931-1936.
26. Sturn A, Quackenbush J, Trajanoski Z. Genesis: cluster analysis of microarray data. *Bioinformatics*. 2002;18(1):207-208.
 27. Mao TK, Chen CZ. Dissecting microRNA-mediated gene regulation and function in T-cell development. *Methods Enzymol*. 2007;427:171-189.
 28. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 2001;25(4):402-408.
 29. Van Os RP, Dethmers-Ausema B, De Haan G. In vitro assays for cobblestone area-forming cells, LTC-IC, and CFU-C. *Methods Mol.Biol.* 2008;430:143-157.
 30. Ploemacher RE, Van der Sluijs JP, Van Beurden CA, Baert MR, Chan PL. Use of limiting-dilution type long-term marrow cultures in frequency analysis of marrow-repopulating and spleen colony-forming hematopoietic stem cells in the mouse. *Blood* 1991;78(10):2527-2533.
 31. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 2005;120(1):15-20.
 32. Popovic R, Riesbeck LE, Velu CS et al. Regulation of mir-196b by MLL and its overexpression by MLL fusions contributes to immortalization. *Blood* 2009;113(14):3314-3322.
 33. Zhan M, Miller CP, Papayannopoulou T, Stamatoyannopoulos G, Song CZ. MicroRNA expression dynamics during murine and human erythroid differentiation. *Exp. Hematol.* 2007;35(7):1015-1025.
 34. Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*. 2005;11(3):241-247.
 35. Harrison DE, Astle CM. Loss of stem cell repopulating ability upon transplantation. Effects of donor age, cell number, and transplantation procedure. *J.Exp.Med.* 1982;156(6):1767-1779.
 36. Dong Q, Meng P, Wang T et al. MicroRNA let-7a inhibits proliferation of human prostate cancer cells in vitro and in vivo by targeting E2F2 and CCND2. *PLoS.One*. 2010;5(4):e10147.
 37. Zhu JW, Field SJ, Gore L et al. E2F1 and E2F2 determine thresholds for antigen-induced T-cell proliferation and suppress tumorigenesis. *Mol.Cell Biol*. 2001;21(24):8547-8564.
 38. Li FX, Zhu JW, Hogan CJ, DeGregori J. Defective gene expression, S phase progression, and maturation during hematopoiesis in E2F1/E2F2 mutant mice. *Mol. Cell Biol*. 2003;23(10):3607-3622.
 39. Zhang L, Flygare J, Wong P, Lim B, Lodish HF. miR-191 regulates mouse erythroblast enucleation by down-regulating Riok3 and Mxi1. *Genes Dev*. 2010
 40. Guo S, Lu J, Schlanger R et al. MicroRNA miR-125a controls hematopoietic stem cell number. *Proc.Natl.Acad.Sci.U.S.A* 2010;107(32):14229-14234.
 41. O'Connell RM, Chaudhuri AA, Rao DS et al. MicroRNAs enriched in hematopoietic stem cells differentially regulate long-term hematopoietic output. *Proc.Natl.Acad.Sci.U.S.A* 2010;107(32):14235-14240.
 42. Bousquet M, Harris MH, Zhou B, Lodish HF. MicroRNA miR-125b causes leukemia. *Proc.Natl.Acad.Sci.U.S.A* 2010
 43. Ooi AG, Sahoo D, Adorno M et al. MicroRNA-125b expands hematopoietic stem cells and enriches for the lymphoid-balanced and lymphoid-biased subsets. *Proc.Natl.Acad.Sci.U.S.A* 2010
 44. Zhao X, Tang Y, Qu B et al. MicroRNA-125a contributes to elevated inflammatory chemokine RANTES levels via targeting KLF13 in systemic lupus erythematosus. *Arthritis Rheum*. 2010;62(11):3425-3435.
 45. Van Zant G, Holland BP, Eldridge PW, Chen JJ. Genotype-restricted growth and aging patterns in hematopoietic stem cell populations of allophenic mice. *J.Exp. Med.* 1990;171(5):1547-1565.

CHAPTER 7

CELLULAR BARCODING TOOL FOR CLONAL ANALYSIS IN THE HEMATOPOIETIC SYSTEM

Alice Gerrits, Brad Dykstra,
Olga J. Kalmykova, Karin Klauke,
Evgenia Verovskaya, Mathilde J.C. Broekhuis,
Gerald de Haan and Leonid V. Bystrykh

*Blood. 2010 Apr 1;115(13):2610-8
Appeared on the cover of Blood*

ABSTRACT

Clonal analysis is important for many areas of hematopoietic stem cell research, including *in vitro* cell expansion, gene therapy, and cancer progression and treatment. A common approach to measure clonality of retrovirally transduced cells is to perform integration site analysis using Southern blotting or polymerase chain reaction-based methods. Although these methods are useful in principle, they generally provide a low-resolution, biased, and incomplete assessment of clonality. To overcome those limitations, we labeled retroviral vectors with random sequence tags or "barcodes". On 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. By coupling the barcoding method to a sequencing-based detection system, we could identify major and minor clones in 2 distinct cell culture systems *in vitro* and in a long-term transplantation setting. In addition, we demonstrate how clonal analysis can be complemented with transgene expression and integration site analysis. This cellular barcoding tool permits a simple, sensitive assessment of clonality, and holds great promise for future gene therapy protocols in humans, and any other applications when clonal tracking is important.

INTRODUCTION

Hematopoiesis involves the tightly coordinated process of blood cell production, and is maintained by a small number of hematopoietic stem cells (HSCs). Resolving the exact number of HSCs that actively contributes to hematopoiesis at any given time and tracking the contribution of individual HSCs to each of the different blood cell lineages is important for a better understanding of both normal and malignant hematopoiesis. Performing such analyses in systems containing multiple HSC-derived clones has remained challenging because under normal circumstances the progeny of genetically identical HSCs are indistinguishable from each other. Therefore, analyzing the behavior of normal HSC clones has been limited to measuring their behavior on an individual basis by limiting dilution¹⁻³ or single purified cell⁴⁻⁶ assays. However, clonal analysis has also been successful in hematologic disorders in which abnormal cells often descend from a single common ancestor, and in which malignant clones could be identified by tracking unique genetic rearrangements, deletions or point mutations (reviewed in Gilliland et al⁷).

The realization that unique genetic mutations could be exploited for clonal analysis, combined with the discovery that retroviral vectors could be used to introduce new genetic material into HSCs,⁸⁻¹⁰ led to the development of an alternative method to perform clonal analysis on hematopoietic cell populations. In recipients that received a transplant with retrovirally transduced HSCs, it was possible to distinguish between different HSC-derived clones by considering retroviral integration sites as unique clonal marks.^{11;12} In this type of analysis, genomic DNA is fragmented with restrictases that cut within the vector and throughout the genome, resulting in fragments that consist of a small part of the vector and adjacent genomic DNA. Different integration sites thereby generate fragments of different lengths that can be detected by Southern blotting.¹³⁻¹⁵ Alternatively, these fragments can be amplified with the use of polymerase chain reaction (PCR)-based approaches in which linear fragment ends are self-ligated (inverse PCR),¹⁶ or in which primer tags are attached to the fragment ends (e.g. ligation-mediated PCR and linear amplification-mediated PCR).^{17;18} Although these retroviral marking experiments did confirm the long-term self-renewal activity and multilineage differentiation potential of HSCs, the methods used have serious shortcomings and could therefore provide only a rough estimation of clonality. Southern blotting relies on restriction digestions; therefore, integrations without an appropriately positioned restriction site cannot be detected. Further, the method lacks sensitivity as well as discriminatory power to count the total number of integrated vectors. Although the PCR-based approaches are more sensitive, they still depend on restriction digestions and are prone to experimental error because of unequal PCR amplification caused by variations in template melting properties and relative differences in fragment length.

To overcome these limitations, we constructed a retroviral plasmid library that consisted of vectors containing a variable random sequence tag or “barcode”. On stable chromosomal integration, this barcode introduces a unique, identifiable, and heritable mark into the genome, allowing the clonal progeny of the host cell to be tracked over time. Here, we couple the concept of cellular barcoding to a sequencing-based detection system, and show the efficacy of the barcoding method to track clonal dynamics in 2 distinct cell culture systems *in vitro* and in a hematopoietic transplantation setting *in vivo*. We also show that this cellular barcoding method can be complemented with transgene expression analysis and integration site analysis, providing additional layers of information.

METHODS

Mice

C57Bl/6 (CD45.2) and DBA/2 mice were purchased from Harlan. C57Bl/6.SJL (CD45.1) mice were bred and all animals were housed at the Central Animal Facility of the University of Groningen. All animal experiments were approved by the Groningen University Animal Care Committee.

Construction of barcoded vector libraries

The SF91 vector was kindly provided by Prof. C. Baum (Hannover Medical School, Hannover, Germany), and the MIEV vector was kindly provided by Prof. C. Jordan (University of Rochester, Rochester, NY). The barcode linker was created by annealing two 5′ phosphorylated primers (forward, 5′-GTACAAGTAANNATCNGATSSAAANNGGTNNAACNNTGTA AACGACGGCCAGTGAG-3′; reverse, 5′-GATCCTCACTGGCCGTCGTTTTACANNGTTNACCNNTTTSSATCNGATNNTTACTT-3′; Biolegio). Primers were dissolved in 0.5x ligation buffer (Fermentas) at a concentration of 100 μM. After heating the mixture for 5 min at 95°C, the primers were allowed to anneal at gradually decreasing temperature. The annealed barcode linker was ligated into the *BsrGI-BamHI* site of the SF91 or MIEV vector (Figure 1A) at equimolar ratio. The resulting vector was transformed into 10-beta competent *Escherichia coli* cells (New England Biolabs) and grown overnight on LB plates supplemented with 50 μg/ml ampicillin (Sigma-Aldrich). Colonies were pooled by flushing plates with LB supplemented with 50 μg/ml ampicillin. After overnight culture, plasmid DNA was extracted using the GenElute HP Plasmid Midiprep Kit (Sigma-Aldrich).

Validation of barcoded vector library

From an SF91 vector library created by combining ~800 bacterial clones, barcode sequences were amplified using primers directed against internal vector sequences (SF91 eGFP forward, 5′-CTGCCCGACAACCACTACCTG-3′; SF91

WPRE reverse: 5'-CCCTAAAAATGTAAATGATTGCCCCACC-3'). The resulting mixture of equal-sized PCR products (i.e. "crude" PCR product) was purified and sequenced by ServiceXS or StarSEQ using a primer directed against eGFP (eGFP forward: 5'-GCGATCACATGGTCCTGCTG-3'). The complexity of this SF91 library was validated by retransforming it into 10-beta competent *E. coli* (New England Biolabs). Single colonies were picked and grown in LB supplemented with 50 µg/ml ampicillin and 30 µg/ml kanamycin (Sigma-Aldrich). Plasmid DNA from each bacterial clone was isolated using the GeneJET Plasmid Miniprep Kit (Fermentas) and sequenced using the same primers as for the "crude" PCR sequencing. Pictogram (<http://genes.mit.edu/pictogram.html>) was used to visualize the combined monoclonal sequence traces.

Cell culture

The 32D (murine myeloid progenitor) cell line was cultured in RPMI-1640 medium with L-glutamine (PAA Laboratories) supplemented with 10% FBS, penicillin and streptomycin (Invitrogen) and 10 ng/mL recombinant murine interleukin-3 (R&D Systems). Bone marrow (BM) cells were isolated from mice 4 days after intraperitoneal injection of 150 mg/kg 5-fluorouracil (Pharmachemie Haarlem), and cultured in StemSpan (StemCell Technologies) supplemented with 10% fetal bovine serum, 300 ng/mL polyethylene glycol-complexed recombinant rat stem cell factor (Amgen), 20 ng/mL rmlL11 (R&D systems), 1 ng/mL Flt3 ligand (Amgen), penicillin and streptomycin.

Retroviral transduction procedure

The 32D cell line and primary BM cells were transduced as previously described.¹⁹ Briefly, Phoenix ecotropic packaging cells were transfected with 1 µg of barcoded SF91 or MIEV vectors. Virus-containing supernatant harvested 24 and 48 hours later was used to transduce 2×10^5 32D cells or 7.5×10^5 BM cells per 3.5 cm well. Transduction efficiencies were determined by flow cytometry (FACSCalibur, Becton Dickinson) and were ~90% for 32D cells and 30-50% for BM cells. 32D cells were transduced with a barcoded SF91 vector library containing ~700 barcodes, BM cells for culture were transduced with a barcoded MIEV vector library containing ~700 barcodes, and BM cells for transplantation studies were transduced with either a low complexity (LC) or high complexity (HC) SF91 barcode library containing ~50 or ~800 barcodes, respectively.

Initiation of polyclonal and monoclonal cultures

Four days after the initial transduction, 2×10^4 eGFP⁺ 32D cells were sorted using a MoFlo flow cytometer (Beckman Coulter). The resulting polyclonal culture was maintained for 5 weeks; extensive barcode analysis was performed at weeks 2 and 5. Monoclonal cultures were initiated by sorting single eGFP⁺ 32D cells into 96-well round-bottom plates and culturing those for ~3 additional weeks; 2 of

these were selected for extensive barcode analysis. Polyclonal DBA/2 BM cultures were initiated with 8×10^5 transduced cells, without sorting for eGFP expression. These cultures were maintained for 5 weeks; extensive barcode analysis was performed for one selected culture at weeks 1 and 5. To generate monoclonal BM cultures, single eGFP⁺Sca-1⁺CD48⁻EPCR⁺ cells from day 7 transduction cultures of C57Bl/6 BM were sorted into 96-well round-bottom plates, and clones reaching population sizes of ~30,000 were selected for further analysis.²⁰

Transplantation and cell purification

LC and HC-barcoded C57Bl/6.SJL BM cells (transduction efficiencies 35% and 50% respectively) were transplanted into lethally irradiated (9,5Gy, IBL 637 ¹³⁷Cs γ -source, CIS Biointernational) C57Bl/6 mice without prior sorting for eGFP expression. Of both the LC- and the HC-barcoded cells 10^6 and 5×10^6 were transplanted into two recipients each. At 8, 17 and 33 weeks after transplantation, blood samples were taken from the retro-orbital plexus. Erythrocytes were lysed in ammonium chloride solution and the remaining cells were stained with APC-conjugated antibody to Gr-1 (clone RB6-8C5) and PE/Cy7-conjugated antibody to CD3 ϵ (clone 145-2C11, both from BD Biosciences Pharmingen). eGFP⁺ granulocytes (Gr-1^{hi}, SSC^{hi}) and T cells (CD3⁺, SSC^{lo}) were sorted using a MoFlo flow cytometer and were stored in RNAlater (QIAGEN) for barcode analysis. One of the mice transplanted with 10^6 LC-barcoded cells (recipient 2) was excluded from further analysis, because the number of eGFP⁺ cells that could be isolated was not sufficient for barcode analysis.

Barcode recovery and identification

Genomic DNA was extracted from cultured 32D and BM cells using the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich). From the purified blood cells, genomic DNA was extracted using the REExtract-N-Amp Tissue PCR Kit (Sigma-Aldrich) followed by a clean-up step using the Nucleospin Plasma XS kit (BioKe, Macherey-Nagel). Total RNA was isolated from cultured 32D cells using the RNeasy Mini kit (QIAGEN) and standard cDNA synthesis (Invitrogen) was performed. Barcode sequences were amplified using primers directed against internal vector sequences (SF91 eGFP forward and SF91 WPRE reverse; MIEV eGFP forward, 5'-CTGCCCCGACAACCACTACCTG-3', and MIEV long terminal repeat reverse, 5'-CCAAACCTACAGGTGGGGTCTTTCATTC-3'). The mixture of equal-sized PCR products (i.e. "crude" PCR product) was purified and sequenced as already described. To obtain individual barcodes for sequencing, the "crude" PCR products were subcloned into pCR4-TOPO-TA (Invitrogen) and transformed into 10-beta competent *E.coli* cells (New England Biolabs).

Binomial distribution modeling

To estimate barcode frequencies we applied a binomial model. The binomial distribution determines the probability $P(k;n,p)$ of observing a particular barcode k times in n sequencing trials, given a probability p of observing this barcode in a single trial. Under the assumption that all barcodes have an equal chance of being selected, the probability p equals the inverse of the barcode complexity B (ie, $p=1/B$) for all barcodes. Assuming independence of the individual barcode detections, the expected frequency of observing some barcode k times within n sequencing trails has been estimated by multiplying the number of barcodes B with $P(k;n,1/B)$. To estimate the barcode complexity B we minimized the mean square distance between the expected and the experimentally observed frequencies.

Integration site analysis

To simultaneously identify barcodes and retroviral integration sites, inverse PCR was performed. 0.1 μg of DNA from monoclonal MIEV-transduced BM cultures was digested with *TatI* (Fermentas), followed by heat inactivation. DNA fragments were then diluted to a concentration of 0.7 ng/ μl and ligated overnight with 10U T4 DNA ligase (Fermentas). After purification of circular fragments (Bioke), a first round of PCR amplification (1-PCR) was initiated with 9 ng of DNA and PCR master mix (Fermentas), under the following conditions: 94°C x 2 min; 30 cycles: 94°C x 1 min, 59°C x 1 min, 72°C x 3.5 min; 72°C for 10 min (forward, 5'-ACCTGTAGGTTTGGCAAGCTAGCTTAAG-3'; reverse, 5'-CAAACCTACAGGTGGGGTCTTTCATTC-3'). A second round of PCR amplification was initiated with 2 μl of 1-PCR reaction, under the following conditions: 94°C x 2 min; 20 cycles: 94°C x 1 min, 58°C x 1 min, 72°C x 3 min; 72°C for 10 min (forward, 5'-GTTTGCATCCGAATCGTGGACTC-3'; reverse, 5'-CCTCACTGGCCGTCGTTTTAC-3'). After electrophoresis, observed fragments were cut from the gel, subcloned in pCR4-TOPO-TA (Invitrogen) and sequenced. Subsequently, retroviral integration sites were identified for 2 monoclonal cultures performing a Mouse BLAT Search on the University of California Santa Cruz Genome Browser (Assembly July 2007, NCBI Build 37).²¹

RESULTS

Construction of barcoded vector library

Random sequence tags or “barcodes” were generated by annealing 2 synthetic oligonucleotides consisting of sets of random nucleotides separated by fixed triplets. The resulting double-stranded linker was then ligated into different types of retroviral vectors (Figure 1A). The fixed triplets in the barcode linker served a dual purpose. First, they enabled the annealing of forward and reverse primers into the barcode linker during initial construction of the barcoded vector

library. Second, the fixed triplets facilitated the analysis of sequencing results by providing an internal standard to evaluate the quality of each sequence trace. The primer binding site was added to the barcode sequence so that barcode-positive and -negative clones could be easily distinguished with the use of PCR. Although in theory more than 4 million ($4^{10} \times 2^2$) possible linker variants could be generated, in practice the number of variants was restricted to the number of bacterial clones generated upon transformation. By combining different numbers of bacterial clones, vector libraries of different complexities (ie, consisting of different numbers of barcodes) were generated. From each resulting vector library, barcode sequences were amplified using primers directed against internal vector sequences, after which the resulting mixture of equal-sized PCR products was sequenced. The resulting “crude” sequence traces suggested that a largely random mixture of sequence tags was generated, because N and S positions were found to be essentially equal in all 4 (A, T, C, and G) or 2 channels (C, and G), respectively (Figure 1A). Using both restriction analysis and PCR, we could not detect unbarcoded vectors, demonstrating that the efficiency of vector barcoding was close to 100% (data not shown).

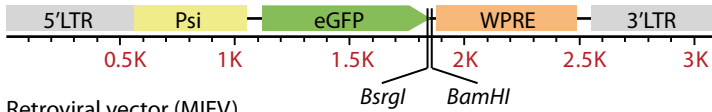
Validation of barcoded vector library

To validate the complexity of the prepared barcode libraries, we transformed a high complexity vector batch (created by combining ~800 bacterial clones) back into *E. coli* and performed monoclonal sequence analysis. Of 88 sequencing trials, 78 represented unique barcodes. By combining the sequence traces of all 88 trials the “crude” PCR sequence trace could be reconstructed (Figure 1B), confirming that the insertion of nucleotides at each variable position was close to random. We next applied a binomial model to approximate the total complexity of the library (for details see “Binomial distribution modeling”), and estimated it to be ~440 (Figure 1C). It should be noted that this prediction of the barcode complexity is only a rough estimate, because the number of sequencing trials was much lower than the complexity of the barcoded library. The accuracy of this prediction becomes progressively higher as the number of sequencing trials approaches or exceeds the number of barcodes in the library.

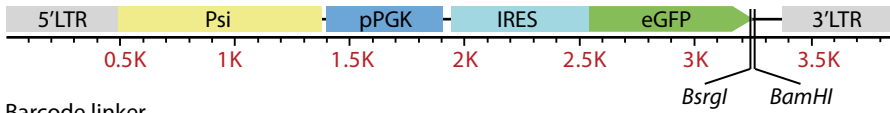
We also assessed the number of nucleotide differences between the 78 unique barcodes identified in the vector library, and found that most barcodes differed

Figure 1. Construction and validation of barcoded vector library. (A) Insertion of the barcode linker into retroviral vectors SF91 and MIEV. The linker contains a variable sequence part that consists of pairs of degenerate nucleotides (N or S) flanked by fixed triplets. The “crude” PCR sequence trace of the resulting vector batch suggests the random insertion of nucleotides at N and S positions. (B) The HC vector library, created by combining ~800 bacterial clones, was re-transformed into *Escherichia coli*. By combining the sequence traces of the 88 resulting clones the “crude” PCR sequence trace could be reconstructed. (C) Distribution plot showing observed and expected barcode frequencies ►

A Retroviral vector (SF91)

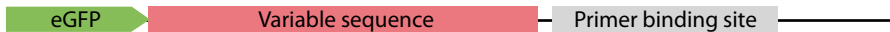


Retroviral vector (MIEV)

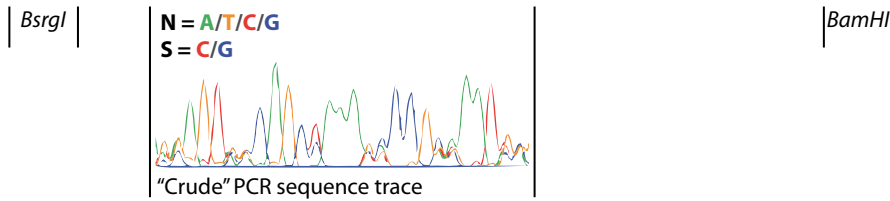


Barcode linker

4.194.304 possible combinations



GTACAAGTAA**NNATC**NNGAT**SSAAANN**GGT**NNAAC**CNN**TG**TAAAACGACGGCCAGTGAG
 TTCATT**NNTAG**NNCT**ASS**T**TNNCC**ANN**TG**NNACATTTTGCTGCCGGTCACTCCTAG



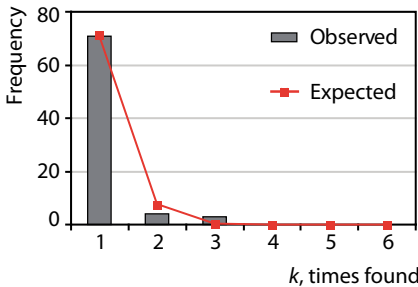
B



88 trials

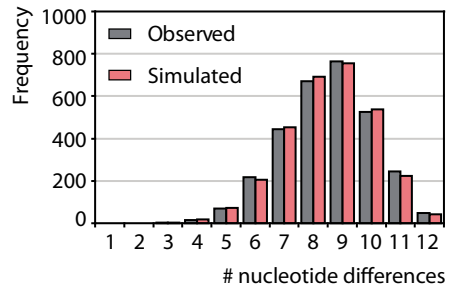
C

Barcode frequency plot
Estimated complexity ~440



D

Barcode comparison plot
78 barcodes



► for the HC vector library; given that 88 sequencing trials were performed. Binomial distribution modeling estimates a barcode complexity of ~440. (D) Distribution plot showing the observed number of nucleotide differences between all 78 barcodes in the HC library by performing pairwise comparisons. Also shown is the distribution of the predicted number of differences based on the simulation of 78 random barcodes (average of 10 simulations). LTR indicates long terminal repeat; Psi, packaging signal; eGFP, enhanced green fluorescent protein; WPRE, woodchuck hepatitis virus post-transcriptional regulatory element; pPGK, phosphoglycerate kinase promoter; IRES, internal ribosome entry site.

in 9 of 12 variable nucleotides. When this was compared to data generated by simulating the generation of 78 random barcodes, the frequency distributions were almost identical (Figure 1D). Together, these data confirmed that the number of bacterial clones that are combined to generate vector libraries can be used as a rough estimate of the complexity of these libraries, and that the generated vector libraries are indeed composed of vectors marked with randomly generated barcodes.

Clonal analysis of polyclonal 32D culture

As a first validation of the cellular barcoding method, we performed clonal analysis on an immortalized cell line that is known to be highly susceptible for retroviral transduction, with high transduction efficiencies and multiple retroviral integrations per cell. We transduced 32D cells with a barcoded vector library, and initiated a polyclonal culture. After 2 weeks of culture, the “crude” PCR sequence trace revealed a complex barcode signature that was comparable to the sequence trace of the barcoded vector library. Subclone sequencing confirmed the high degree of polyclonality in this culture, because 77 unique barcodes were identified of 105 sequencing trials. After 3 additional weeks of culture with semiweekly passaging, 68 unique barcodes were identified of 95 sequencing trials (Figure 2A). By applying a binomial model, we were able to estimate that the complexity of this culture decreased slightly from ~210 at week 2 to ~160 at week 5 (Figure 2B). Nevertheless, as would be expected for an immortalized cell line, the barcoded culture largely retained its polyclonal nature over time.

Enumeration of integration sites in 32D monoclonal cultures

We also initiated monoclonal cultures with single transduced 32D cells. Genomic DNA was isolated from 2 selected monoclonal cultures, and barcode sequences were amplified. The “crude” PCR sequence traces of these 2 cultures revealed dissimilar levels of complexity, although both were noticeably simpler than the polyclonal culture (Figure 2C). Subclone sequencing was then performed to identify the underlying barcodes. The monoclonal culture with the simpler “crude” PCR signature (monoclonal I) was found to contain 2 different barcodes whereas the more complex (monoclonal II) was found to contain 9 different barcodes (Figure 2D). Because these cultures were initiated with single cells, each identified barcode represents a unique retroviral insert. It could be determined with confidence that the barcodes detected represented the actual number of integrations, because the probabilities of having missed a third integration site in monoclonal I or a tenth integration in monoclonal II were only 0.035 and 0.022, respectively. Compared to traditional methods for integration site analysis, the barcoding method can determine the actual number of integrations in monoclonal populations with greater confidence. Overall, the “crude” PCR signatures proved to be a good representation of the actual underlying barcodes, because by

combining the individual barcode sequence traces the “crude” PCR sequence trace could be reconstructed (Figure 2D).

Transgene expression analysis in 32D monoclonal

Because the barcode tag is positioned immediately downstream of the eGFP coding sequence, each eGFP mRNA molecule contains the same barcode as the DNA from which it was transcribed. As such, the presence of barcodes at the RNA level permits the direct analysis of clone-dependent and/or integration site-dependent eGFP expression. To demonstrate this, the proportion of eGFP expression corresponding to each of the integrated vectors in the 2 monoclonal was measured by isolating RNA and amplifying the expressed barcode sequences. This transgene expression analysis showed that in both monoclonal eGFP molecules could be detected that were driven from each of the different integration sites (Figure 2E). Although 1 of 9 barcodes from monoclonal II (marked by an asterisk) appeared to be overrepresented on the RNA level, this finding did not reach statistical significance ($p=0.066$), in part because the number of subclones sequenced was relatively small compared with the number of integrations. Nevertheless, if a sufficient number of subclones are sequenced, the barcoding method will be able to identify clones and/or integrations that dominate expression within mixed cell populations.

7

Clonal analysis of primary bone marrow cell culture

To further validate the cellular barcoding method, we performed clonal analysis on primary BM cultures that are less susceptible to retroviral transduction and would probably exhibit decreased clonal complexity over time. BM cells were transduced with a barcoded vector library, and the barcode compositions of the cultures were measured at 2 different points in time. One culture was selected for barcode analysis, and its clonal complexity was measured over time (Figure 3A). After 1 week of culture, the estimated barcode complexity was ~110, and the observed and expected barcode frequency distributions were almost identical (Figure 3B, left panel). In contrast, after 4 additional weeks of culture one barcode was clearly found to dominate in the culture, being found in 36 of 50 sequencing trials. When this obviously overrepresented barcode was removed from the observed frequencies, the estimated complexity of the remaining culture was ~20, assuming that the remaining barcodes were equally distributed (Figure 3B, right panel).

Integration site identification in bone marrow monoclonal

To show that clonal analysis can be complemented with integration site analysis, we purified single MIEV-transduced BM progenitor cells by fluorescence-activated cell sorting and initiated monoclonal cultures. Genomic DNA was isolated from several of these cultures, and inverse PCR was performed to simultaneously

identify integration sites and corresponding barcodes of the inserted retroviral vector (2 examples are shown in Figure 3C).

Clonal analysis of repopulation dynamics *in vivo*

Having established the feasibility of using cellular barcoding as a tool for clonal analysis *in vitro*, we next used this method to track hematopoietic repopulation dynamics *in vivo* following transplantation of barcoded HSCs. First, to track general trends in hematopoietic reconstitution, we transduced BM cells with a relatively small number of barcodes (LC, ~50) and transplanted 5 million cells. Second, to directly track the clonal behavior of individual HSCs, we transduced BM cells with a relatively high number of barcodes (HC, ~800) and transplanted only 1 million cells. At 8, 17 and 33 weeks after transplantation, eGFP⁺ granulocytes and T cells were sorted from the peripheral blood and barcode analysis was performed by sequencing 19 to 30 subclones from each blood cell type at each time point (Figure 4).

In both mice, the barcode signature was dominated by relatively few barcodes at all time points, suggesting that within the eGFP⁺ population only a few stem/progenitor cells were the main contributors of granulocytes and T cells. It should be noted that the LC-barcoded mouse probably contains multiple HSC-derived clones marked with the same barcode, and so it is likely that the actual number of clones in this mouse exceeds the number of clones in the HC-barcoded mouse.

In both recipients, the barcode signatures of the T cells isolated at 8 weeks after transplantation differed substantially from the signatures of T cells and granulocytes at all other time points. This might be explained by the long life

Figure 2. Barcode analysis of polyclonal and monoclonal 32D cell cultures. (A) Tracking the complexity of a polyclonal 32D cell culture over time. Subclone sequencing revealed 77 unique barcodes of 105 sequencing trials after 2 weeks of culture, and 68 of 95 after 5 weeks of culture. White bars represent barcodes that were identified at 1 time point only, and colored/patterned bars represent barcodes that were found at both time points. (B) Distribution plots showing the observed and expected barcode frequencies for the 2 time points. Binomial distribution modeling predicts that the barcode complexity in this culture remained complex over time. (C) “Crude” PCR sequence traces of 2 monoclonal cultures provide an estimate of the total number of retroviral integrations per cell. The sequence trace for monoclonal I shows that all variable positions in the barcode are restricted to 1 or 2 nucleotides, suggesting the presence of only 2 integration sites. The sequence trace for monoclonal II is more complex, but 1 nucleotide is missing at several of the variable positions. (D) Subclone sequencing identified the unique barcodes that underlie the “crude” PCR sequence traces. Different barcodes are represented by different colors. Monoclonal I contains 2 integrated retroviral vectors per cell, whereas monoclonal II contains 9 integrated retroviral vectors per cell. Also shown are the actual barcode sequences identified for both monoclonals. The sequence consensus pictograms show how the “crude” PCR sequence traces in panel B can be reconstructed by combining all identified barcodes *in silico*. (E) Gene expression analysis identified barcodes that were present at the RNA level. One of 9 barcodes in monoclonal II (marked with an asterisk) appears to be overrepresented on the RNA level, but this finding does not reach statistical significance. Note that identical barcodes in panels D and E of this figure are represented using matching colors.

A 32D polyclonal culture
Subclone sequencing

wk 2



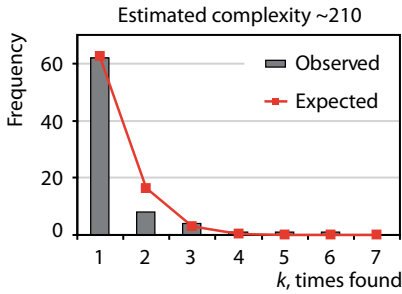
77 unique / 105 trials

wk 5

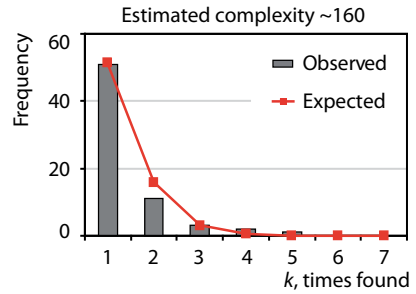


68 unique / 95 trials

B wk 2



wk 5



C 32D monoclonal cultures
retroviral inserts
Crude PCR sequencing

Monoclone I: DNA



Monoclone II: DNA



D 32D monoclonal cultures
retroviral inserts
Subclone sequencing

I: DNA 2 unique / 11 trials



II: DNA 9 unique / 58 trials



E 32D monoclonal cultures
Expression from inserts
Subclone sequencing

I: RNA 2 unique / 12 trials

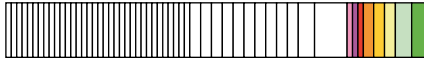


II: RNA 9 unique / 27 trials



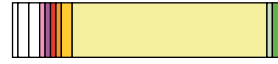
A BM polyclonal culture
Subclone sequencing

wk 1



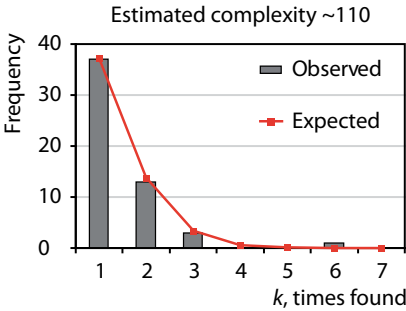
54 unique / 78 trials

wk 5

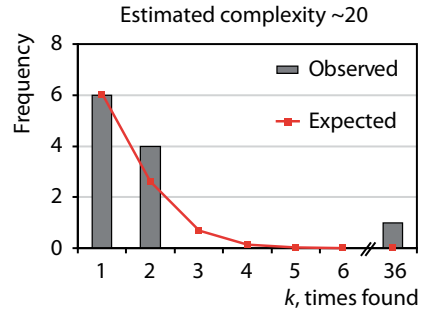


11 unique / 50 trials

B wk 1



wk 5



C BM monoclonal cultures
Integration site ID
Inverse PCR & sequencing

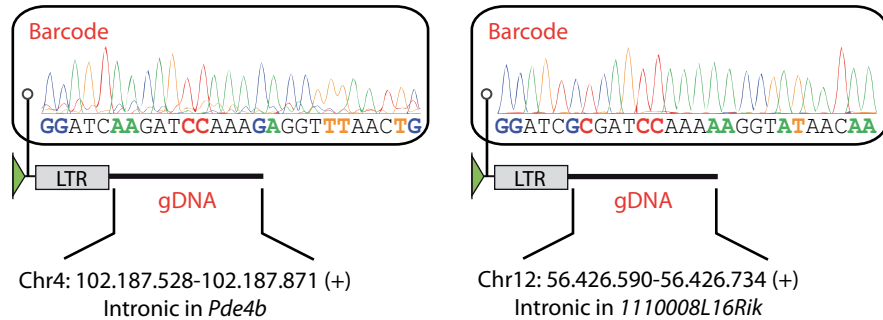


Figure 3. Barcode analysis of primary bone marrow cell cultures. (A) Tracking the clonal complexity of a primary BM culture over time. Subclone sequencing reveals 54 unique barcodes of 78 sequencing trials after 1 week of culture, and only 11 of 50 after five weeks of culture. White bars represent barcodes that were identified at 1 time point only, and colored bars represent barcodes that were found at both time points. (B) Distribution plots showing the observed and expected barcode frequencies for the 2 time points. Binomial distribution modeling predicts that the barcode complexity in the culture markedly decreased over time. Note that the model for week 5 excludes the outlier that was found 36 times and is therefore calculated for 14 trials only. (C) Inverse PCR identifies the integration site and the corresponding barcode for 2 monoclonal cultures initiated with single transduced BM progenitor cells. LTR, long terminal repeat.

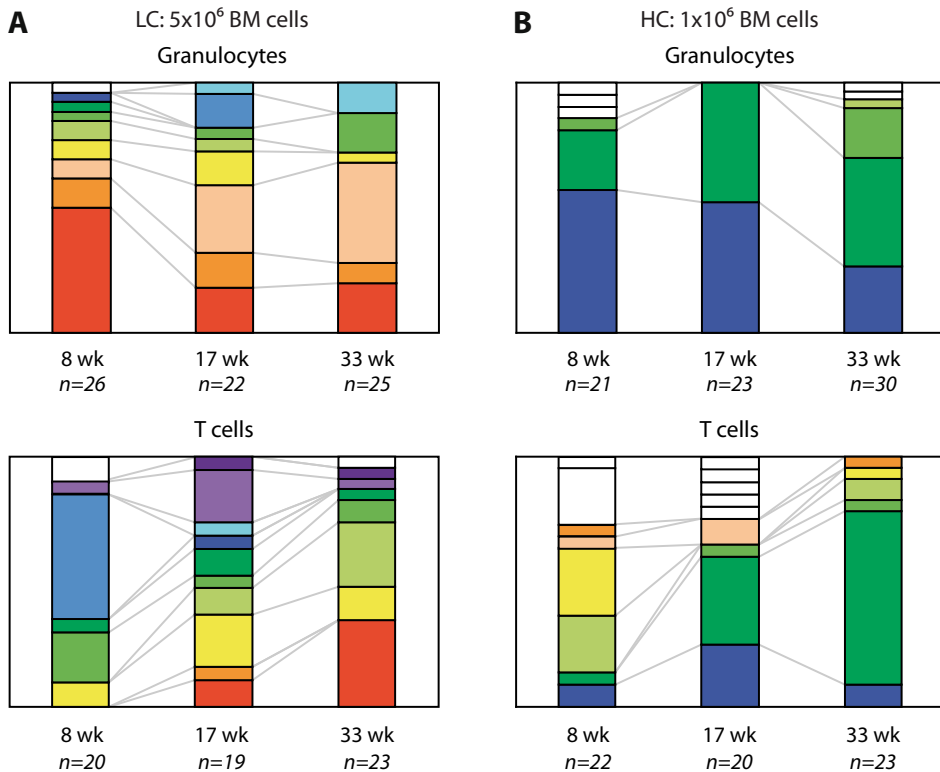


Figure 4. Clonal dynamics of *in vivo* hematopoiesis. Clonal composition of peripheral blood granulocytes and T cells over time in one recipient that received a transplant with 5×10^6 LC-barcoded BM cells (A) and a second recipient that received a transplant with 10^6 HC-barcoded cells (B). White bars represent barcodes that were found in only 1 cell type and time point, and colored bars represent barcodes that were found in both cell types and/or multiple time points.

span of T cells relative to granulocytes. T-lineage cells produced in the first few weeks by transduced multipotent or lineage-restricted progenitors might still result in T cells to be present in the blood at 8 weeks after transplantation. In contrast, most or all multipotent or myeloid-restricted progenitors targeted during the transduction procedure would not continue producing peripheral blood granulocytes for 8 weeks after transplantation. In addition, some barcodes were present in granulocytes at early time points, but appeared in T cells only at later time points. This might also reflect the difference in life span of both cell types, in combination with the difference in time required for these cells to undergo maturation and migrate to the peripheral blood.

Whereas some barcodes marked both granulocytes and T cells, others were identified in only 1 of the 2 cell types. It is possible that these barcodes may have distinguished multi-potent HSCs that contributed to both myelopoiesis and

lymphopoiesis from those that were myeloid-biased or lymphoid-biased in their lineage contributions.

Combined, these data suggest that cellular barcoding can be used for the simultaneous tracking of clones contributing to hematopoiesis. Depending on the complexity of the vector library and the number of subclones sequenced, the resolution of this technique could be increased to track minor clones with more confidence.

Clonal analysis of long-term HSC activity *in vivo*

Next, we used the barcoding method to measure the relative activity of individual HSCs in a polyclonal BM transplantation setting. As a surrogate measurement for stem cell activity, we analyzed the barcodes present in granulocytes, which have a short life span and must therefore be constantly replenished by the currently active HSC clone(s). eGFP⁺ granulocytes were isolated at 33 weeks after transplantation from the peripheral blood of 3 recipients of LC-barcoded cells and 4 recipients of HC-barcoded cells, and 33 to 54 barcodes were sequenced from each sample (Figure 5A). In all recipients at least 1 major clone could be identified that was represented at a level higher than would be expected by chance because of random sampling error from an equally distributed population. Conversely, at least 1 and up to 8 barcodes were detected only once. This suggests that long-term hematopoiesis is for the most part maintained by a few HSC-derived clones, but that underneath these major clones a whole spectrum of minor clones may exist. Indeed, the fact that several barcodes were only seen once suggests that the number of identified minor clones is likely to be an underestimation of the actual number of minor clones in the population. By increasing the depth of sequencing a more accurate snapshot of the HSC compartment can be taken.

As would be expected, a much higher proportion of barcodes was found to be shared between the LC-barcoded mice (5 of 15 barcodes shared between 3 mice) than the HC-barcoded mice (1 of 37 barcodes shared between 4 mice; Figure 5B). This indicates that redundantly marked HSCs were indeed transplanted into the LC-barcoded mice, whereas the complexity of the HC library was sufficient to mark almost all HSCs uniquely.

DISCUSSION

In this study, we demonstrate that cellular barcoding can be used as a powerful tool for clonal analysis in the hematopoietic system. This method overcomes many of the limitations of the Southern blot- and PCR-based techniques that were used for clonal analyses in the past, because the sequencing-based barcode detection system does not depend on restriction digestion or probe hybridization, and is less prone to unequal PCR amplification. Therefore, we can identify with greater confidence both major and minor clones within complex cell populations *in vitro* and *in vivo*.

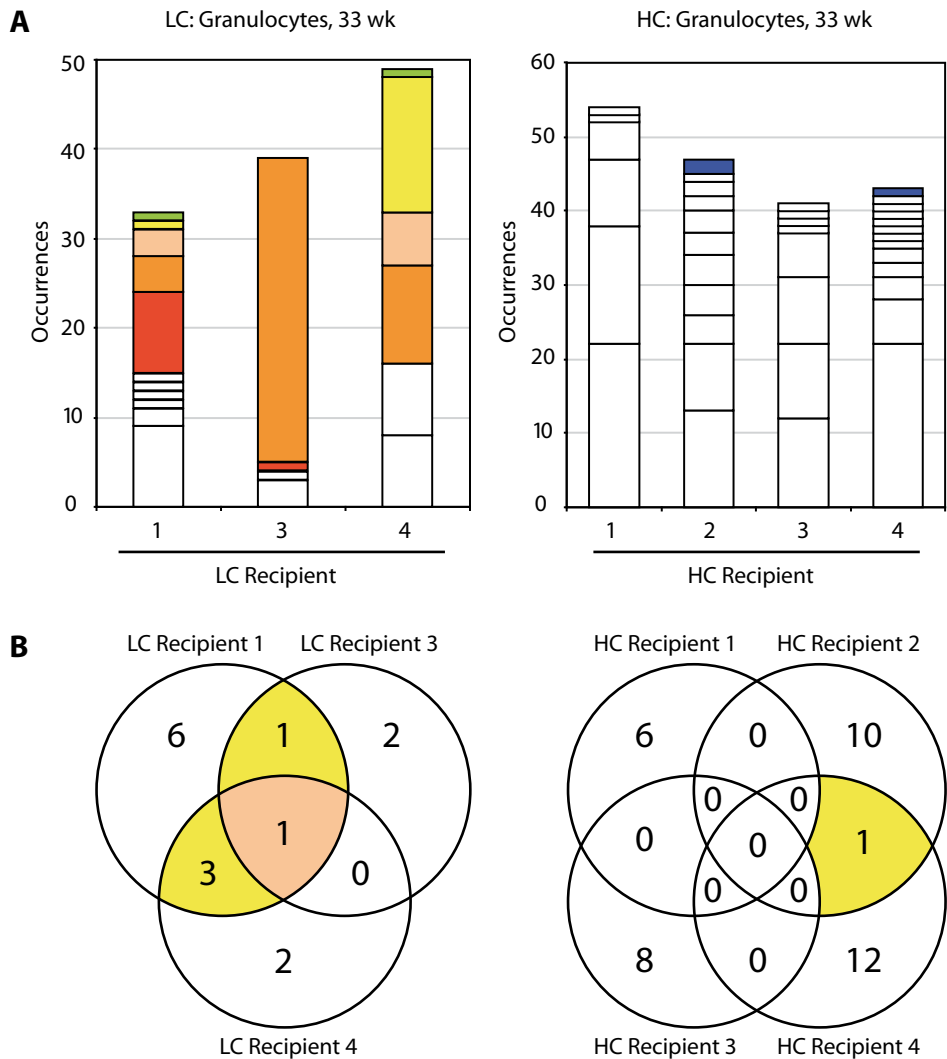


Figure 5. Barcode profiles of long-term HSC activity. (A) Clonal composition of granulocytes at 33 weeks after transplantation in 3 recipients of LC-barcoded cells and 4 recipients of HC-barcoded cells. White bars represent barcodes that were identified in 1 recipient only, and colored bars represent barcodes that were shared between recipients. (B) Venn diagrams showing the degree of redundant barcoding between multiple recipients receiving the same starting cells. Recipients 1 and 2 received a transplant with 1 million cells, recipients 3 and 4 received a transplant with 5 million cells. Note that a portion of the barcodes shown for LC recipient 4 and HC recipient 1 are also presented in Figure 4.

The data presented here provide important insight into the spectrum of active stem cell clones in a transplantation setting. In particular, our observations suggest that hematopoiesis at any given time appears to be dominated by a few HSC-derived clones, with additional contributions from a number of minor clones. Although the exact number of minor clones that are contributing at any given time remains undefined because of a lack of sequencing depth, it is still interesting to speculate about their biological identity. Although their behavior could be due to simple stochastic variation, they might represent one or more functionally distinct HSC subtypes. A further implication is that current functional definitions for HSCs may need to be reconsidered, given that the output of some minor clones might not exceed the commonly used thresholds for repopulation.

Cellular barcoding can also be applied to help resolve other outstanding questions in HSC biology. For example, the concept of lineage-biased HSCs has been suggested,^{6,22} but is still controversial. By tracking barcodes in various cell types, lineage-biased clones could be directly confirmed in a competitive polyclonal setting. Similarly, the concept of dormant or hibernating HSCs has been suggested by several groups.²³⁻²⁵ Cellular barcoding has the potential to measure the differentiated cell output from individual HSCs at different points in time, thereby enabling the identification of dormant or hibernating HSCs. It would be interesting to continue tracking these same clones after subjecting them to serial transplantation or other hematopoietic insult, to determine whether or not their properties can be reset or modified. In addition, different subpopulations of the primitive hematopoietic hierarchy could be purified by fluorescence-activated cell sorting from a recipient that received a transplant with barcoded HSCs, and the clonal relationships between the cell compartments could be analyzed. Compared to previous studies using single-cell transplants, the amount of information that could be gleaned from one transplant recipient would be considerable. As well, such data would be useful to corroborate or further inform long-standing theoretical concepts and quantitative models on the clonal dynamics of HSCs.^{26,27} Collectively, these applications of the cellular barcoding method have the potential to greatly increase our understanding of clonal dynamics in a BM transplantation setting.

In 2 recent reports, cellular barcoding was coupled to a high-throughput microarray-based detection system to study lineage relationships between T-cell subsets²⁸ and T-cell recruitment and expansion in response to infection.²⁹ Although this approach does allow clonal analysis to be performed on the population level, it requires a labor-intensive selection of barcodes and custom-made microarrays. In contrast, the approach described in this study can be more easily implemented. At the same time, a limitation of the approach presented here is that the resolution by which minor clones in a complex mixture can be identified is restricted by the number of subclones sequenced. To overcome this limitation, we are in the process of coupling the method to a high-throughput sequencing-based detection

system, which will permit the simultaneous measurement of all uniquely marked clones at the population level.

Although retroviral vectors have already been used for decades as tools to introduce new genetic material into HSCs⁹⁻¹⁰ it has become clear only recently that insertional mutagenesis from the retroviral integrations themselves can trigger clonal expansion of HSCs.^{18,30} Particularly in this context it is of great added value to barcode retroviral vectors in such a way that clonal analysis can be complemented with transgene expression and integration site analysis. To streamline this process in future experiments, we have now designed a next-generation vector in which the barcode is placed closer to the long terminal repeat. By first enriching for retroviral inserts, followed by high-throughput sequencing, we anticipate that integration sites and their corresponding barcodes can be recovered simultaneously. At the same time, barcoded cDNA from the same cell populations could be sequenced. By normalizing the barcodes found in the cDNA to those found in the DNA, expression could be determined for each integration site in the entire population. Such multi-level analysis would be impossible using unbarcoded vectors.

Recently, a successful HSC gene therapy trial involving 2 patients with X-linked adrenoleukodystrophy was reported,³¹ which has renewed interest and optimism in the gene therapy field.^{32,33} It is exciting to consider how future applications of clinical gene therapy protocols would benefit from the use of barcoded vectors because this would allow a simple and sensitive detection of clonal dominance. For this purpose, the complexity of the library need not be particularly high, because the events triggering clonal dominance are rare. For example, a barcode-to-HSC ratio of 10:1 would already result in an average of ~90% uniquely marked cells (assuming equal representation of barcodes and disregarding the possibility of multiple integrations per HSC). Increasing this ratio to 20:1 or 40:1 would only result in an additional 5% or 7% uniquely marked cells, respectively. In the adrenoleukodystrophy gene therapy trial, patients were transplanted with ~250 transduced HSCs (assuming 25 kg of body mass, HSC frequency of 1 in 10⁵ transfused cells, and 17% gene transfer efficiency into HSCs). A barcode library with a complexity of ~2500, which could be generated within a week in a regular laboratory, would uniquely mark ~90% of the transduced HSCs in these patients, with the remaining HSCs sharing a barcode with only 1 or 2 others. This would be easily sufficient to detect clonal dominance, even in its early stages. Because of the simple and flexible nature of the barcoding method, it is applicable to other viral (or even non-viral) gene delivery approaches, and to systems other than hematopoietic. Therefore, we advocate the use of barcoded vectors in all future clinical gene therapy protocols.

ACKNOWLEDGMENTS

We thank Selvi Durmus, Martha Ritsema, Ellen Weersing, and Jaring Schreuder for technical assistance; Geert Mesander and Henk Moes for assistance in cell sorting; and Verena Zuber and Tilo Buschmann for statistical advice. This work was supported by the Dutch Cancer Society (RUG2007-3729); the Netherlands Genomics Initiative (Horizon, 050-71-055); the Netherlands Organization for Scientific Research (Top Talent, 021-002-084 to E.V.; VENI, 916-86-009 to B.D.; and VICI, 918-76-601 to G.d.H.); and by the European Community (Marie Curie RTN EUrythron, MRTN-CT-2004-005499; and EuroSystem, 200720).

REFERENCES

1. Smith LG, Weissman IL, Heimfeld S. Clonal analysis of hematopoietic stem-cell differentiation in vivo. *Proc.Natl.Acad. Sci.U.S.A.* 1991;88(7):2788-2792.
2. Zhong RK, Astle CM, Harrison DE. Distinct developmental patterns of short-term and long-term functioning lymphoid and myeloid precursors defined by competitive limiting dilution analysis in vivo. *J.Immunol.* 1996;157(1):138-145.
3. Cho RH, Muller-Sieburg CE. High frequency of long-term culture-initiating cells retain in vivo repopulation and self-renewal capacity. *Exp.Hematol.* 2000;28(9):1080-1086.
4. Takano H, Ema H, Sudo K, Nakauchi H. Asymmetric division and lineage commitment at the level of hematopoietic stem cells: inference from differentiation in daughter cell and granddaughter cell pairs. *J.Exp.Med.* 2004;199(3):295-302.
5. Ema H, Sudo K, Seita J et al. Quantification of self-renewal capacity in single hematopoietic stem cells from normal and Lnk-deficient mice. *Dev.Cell.* 2005;8(6):907-914.
6. Dykstra B, Kent D, Bowie M et al. Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell.* 2007;1(2):218-229.
7. Gilliland DG, Blanchard KL, Bunn HF. Clonality in acquired hematologic disorders. *Annu.Rev.Med.* 1991;42:491-506.
8. Joyner A, Keller G, Phillips RA, Bernstein A. Retrovirus transfer of a bacterial gene into mouse haematopoietic progenitor cells. *Nature.* 1983;305(5934):556-558.
9. Williams DA, Lemischka IR, Nathan DG, Mulligan RC. Introduction of new genetic material into pluripotent haematopoietic stem cells of the mouse. *Nature.* 1984;310(5977):476-480.
10. Miller AD, Eckner RJ, Jolly DJ, Friedmann T, Verma IM. Expression of a retrovirus encoding human HPRT in mice. *Science.* 1984;225(4662):630-632.
11. Capel B, Hawley RG, Mintz B. Long- and short-lived murine hematopoietic stem cell clones individually identified with retroviral integration markers. *Blood.* 1990;75(12):2267-2270.
12. Jordan CT, Lemischka IR. Clonal and systemic analysis of long-term hematopoiesis in the mouse. *Genes Dev.* 1990;4(2):220-232.
13. Dick JE, Magli MC, Huszar D, Phillips RA, Bernstein A. Introduction of a selectable gene into primitive stem cells capable of long-term reconstitution of the hemopoietic system of W/Wv mice. *Cell.* 1985;42(1):71-79.
14. Keller G, Paige C, Gilboa E, Wagner EF. Expression of a foreign gene in myeloid and lymphoid cells derived from multipotent haematopoietic precursors. *Nature.* 1985;318(6042):149-154.
15. Lemischka IR, Raulet DH, Mulligan RC. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell.* 1986;45(6):917-927.
16. Nolta JA, Dao MA, Wells S, Smogorzewska EM, Kohn DB. Transduction of pluripotent human hematopoietic stem cells demonstrated by clonal analysis after engraftment in immune-deficient mice. *Proc.Natl.Acad. Sci.U.S.A.* 1996;93(6):2414-2419.
17. Schmidt M, Hoffmann G, Wissler M et al. Detection and direct genomic sequencing of multiple rare unknown flanking DNA in

- highly complex samples. *Hum. Gene Ther.* 2001;12(7):743-749.
18. Kustikova O, Fehse B, Modlich U et al. Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. *Science.* 2005;308(5725):1171-1174.
 19. Kamminga LM, Bystrykh LV, De Boer A et al. The Polycomb group gene Ezh2 prevents hematopoietic stem cell exhaustion. *Blood.* 2006;107(5):2170-2179.
 20. Dykstra B, Schreuder J, Bystrykh L, De Haan G. Flow cytometric purification of primitive hematopoietic progenitors from bone marrow transduction cultures permits clonal analysis of retroviral integrations. Paper presented at Annual Meeting of the Society for Hematology and Stem Cells (ISEH). July 9-12, 2008. Boston, MA.
 21. Kent, J. Mouse BLAT Search. <http://genome.ucsc.edu/cgi-bin/hgBlat>. Accessed June 9, 2009.
 22. Muller-Sieburg CE, Cho RH, Thoman M, Adkins B, Sieburg HB. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood.* 2002;100(4):1302-1309.
 23. Yamazaki S, Iwama A, Takayanagi S et al. Cytokine signals modulated via lipid rafts mimic niche signals and induce hibernation in hematopoietic stem cells. *EMBO J.* 2006;25(15):3515-3523.
 24. Wilson A, Laurenti E, Oser G et al. Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell.* 2008;135(6):1118-1129.
 25. Foudi A, Hochedlinger K, Van Buren D et al. Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. *Nat. Biotechnol.* 2009;27(1):84-90.
 26. Roeder I, Horn K, Sieburg HB et al. Characterization and quantification of clonal heterogeneity among hematopoietic stem cells: a model-based approach. *Blood.* 2008;112(13):4874-4883.
 27. Roeder I, Loeffler M. A novel dynamic model of hematopoietic stem cell organization based on the concept of within-tissue plasticity. *Exp. Hematol.* 2002;30(8):853-861.
 28. Schepers K, Swart E, Van Heijst JW et al. Dissecting T cell lineage relationships by cellular barcoding. *J. Exp. Med.* 2008;205(10):2309-2318.
 29. Van Heijst JW, Gerlach C, Swart E et al. Recruitment of antigen-specific CD8+ T cells in response to infection is markedly efficient. *Science.* 2009;325(5945):1265-1269.
 30. Hacein-Bey-Abina S, Von Kalle C, Schmidt M et al. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science.* 2003;302(5644):415-419.
 31. Cartier N, Hacein-Bey-Abina S, Bartholomae CC et al. Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science.* 2009;326(5954):818-823.
 32. Naldini L. Medicine. A comeback for gene therapy. *Science.* 2009;326(5954):805-806.
 33. Gene therapy deserves a fresh chance. *Nature.* 2009;461(7268):1173.

CHAPTER
SUMMARIZING DISCUSSION



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 *Zfp521*. 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

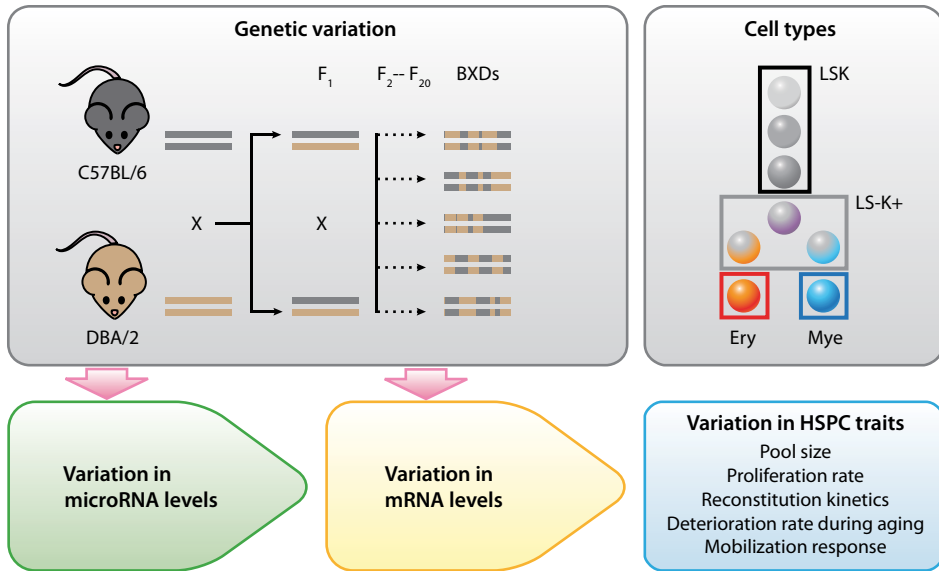
expression study on four hematopoietic cell types isolated from the B6 and D2 parental mouse strains. We identified 131 microRNAs that were differentially expressed between cell types and 15 that were differentially expressed between mouse strains. Of special interest was an evolutionary conserved microRNA 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



8

Induced genetic variation

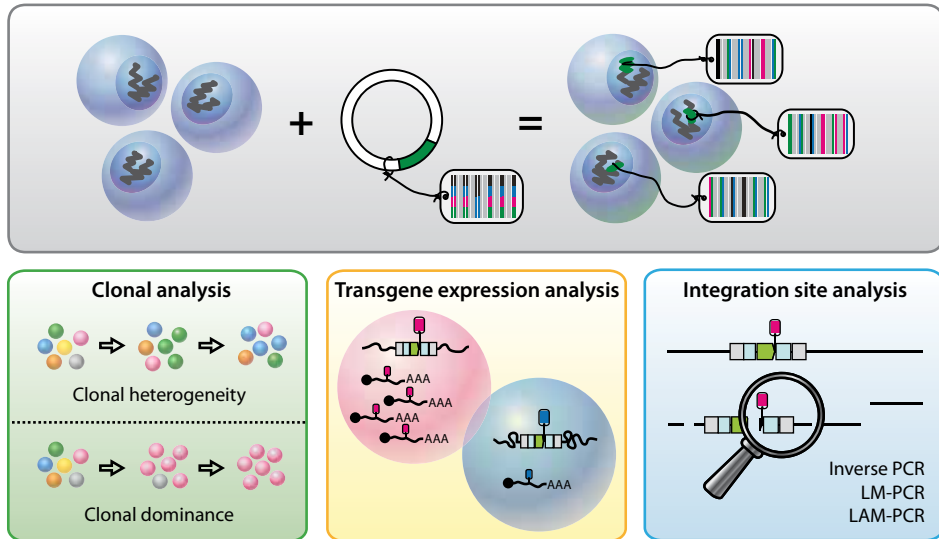


Figure 1. Illustrated summary of this thesis.

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.

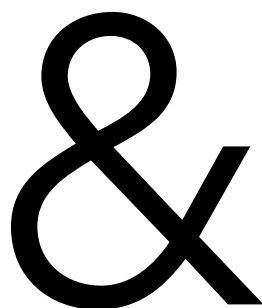
Natural and induced genetic variation united

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.

APPENDICES



CONTRIBUTING AUTHORS

NEDERLANDSE SAMENVATTING
(VOOR NIET-INGEWIJDEN)

DANKWOORD

BIOGRAFIE

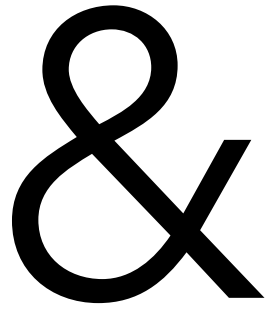
BIOGRAPHY

LIST OF PUBLICATIONS

ATTENDED MEETINGS

FUNDING AND AWARDS

CONTRIBUTING AUTHORS

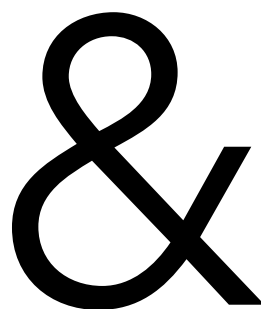


Ausema, Albertina ¹	Jansen, Ritsert C ⁴	Ritsema, Martha ¹
Bot, Jan ^{2,3}	Kalmykova, Olga J ¹	Tesson, Bruno M ⁴
Breitling, Rainer ⁴	Klauke, Karin ¹	Verovskaya, Evgenia ¹
Broekhuis, Mathilde JC ¹	Li, Yang ⁴	Walasek, Marta A ¹
Bystrykh, Leonid V ¹	Olthof, Sandra ¹	Wang, Xusheng ^{7,8}
Dontje, Bert ¹	Otten, Marcel ¹	Weersing, Ellen ¹
Dykstra, Brad ¹	Reinders, Marcel ^{2,3}	Wessels, Lodewyk ^{3,5}
Göttgens, Berthold ⁶	De Ridder, Jeroen ^{2,3,5}	Zwart, Erik ¹
De Haan, Gerald ¹		

1. Department of Cell Biology, Section Stem Cell Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlandsⁱ
2. Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
3. Netherlands Bioinformatics Center, Nijmegen, The Netherlands
4. Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Haren, The Netherlands
5. Bioinformatics and Statistics, Department of Molecular Biology, Netherlands Cancer Institute, Amsterdam, The Netherlands
6. Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, United Kingdom
7. Department of Anatomy and Neurobiology, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America
8. Institute of Bioinformatics, Zhejiang University, Hangzhou, China

ⁱ Since 2011 part of the European Research Institute on the Biology of Aging (ERIBA), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

NEDERLANDSE SAMENVATTING
(VOOR NIET-INGEWIJDEN)



Bloed is één van de meest bestudeerde weefsels in het menselijk lichaam en bestaat uit cellen en plasma. De verschillende typen cellen in het bloed vervullen uiteenlopende functies. Zo zijn rode bloedcellen verantwoordelijk voor het zuurstoftransport in het lichaam, witte bloedcellen voor de afweer en bloedplaatjes voor de bloedstolling. Omdat veel bloedcellen een beperkte levensduur hebben moeten ze continu vervangen worden. Hiervoor zijn bloedvormende stamcellen verantwoordelijk.

Bloedvormende stamcellen kenmerken zich door een unieke combinatie van twee specifieke eigenschappen: zelfvernieuwing en differentiatie. Als een stamcel zich deelt ontstaan er twee dochtercellen, die elk voor een keuze staan. Ze kunnen een stamcel blijven (zelfvernieuwing), of zich ontwikkelen tot een gespecialiseerde bloedcel (differentiatie). Het is belangrijk dat er, gemiddeld genomen, een goede balans bestaat tussen zelfvernieuwing en differentiatie. Als de balans verschuift naar teveel zelfvernieuwing zal de stamcel populatie zich uitbreiden, wat op de lange termijn zal leiden tot leukemie. Echter, als de balans verschuift naar teveel differentiatie zal de stamcel populatie uitputten, wat op de lange termijn zal leiden tot een tekort aan bloedcellen.

Stamcellen moeten continu kiezen: inactief blijven, of juist delen. Als ze gaan delen moeten ze vervolgens ook nog kiezen tussen zelfvernieuwing of differentiatie. De keuzes die stamcellen maken worden beïnvloed door signalen van binnen en van buiten de cel. Voor een beter begrip van de normale – maar ook de abnormale – bloedvorming is het van belang om uit te zoeken hoe die keuzes precies gemaakt worden. Die informatie zou ons mogelijk ook in staat stellen om de keuzes van stamcellen te beïnvloeden.

In dit proefschrift is de bloedvorming bestudeerd door gebruik te maken van genetische variatie. Dit betekent niets anders dan variaties in het DNA, de drager van erfelijke informatie. Op het DNA bevinden zich ongeveer 30.000 genen, die allemaal via RNA (een soort boodschappers van het DNA) vertaald kunnen worden in één of meerdere eiwitten. DNA en RNA bestaan uit vier verschillende bouwstenen. Elke drie opeenvolgende bouwstenen vormen een soort drieletter 'woord' dat bepaalt welk aminozuur in een eiwit wordt ingebouwd. Een fout in de bouwstenen van het DNA kan dus leiden tot een fout eiwit, waardoor er problemen kunnen ontstaan. Hoewel vrijwel alle cellen in ons lichaam hetzelfde DNA bevatten en dus dezelfde genen, vervullen niet alle cellen dezelfde functie. Dat komt omdat in elk van die verschillende celtypen een andere combinatie van RNA en (dus) eiwitten tot expressie komt.

Natuurlijke genetische variatie

Ieder mens heeft een andere DNA code, een soort unieke genetische vingerafdruk. Dit noemen we natuurlijke genetische variatie. Deze variatie zorgt ervoor dat we allemaal van elkaar verschillen. Denk bijvoorbeeld aan verschillen in lichaamslengte, lichaamsbouw, sportiviteit, muzikaliteit, intelligentie en gedrag, maar ook aan verschillen in aanleg om ziektes zoals kanker te ontwikkelen. Het



betreft hier allemaal complexe eigenschappen die bepaald worden door een samenspel van meerdere genen en de omgeving. In dit proefschrift is natuurlijke genetische variatie gebruikt om de (complexe) bloedvorming in muizen te bestuderen.

In hoofdstuk 3 werden vier verschillende typen bloedcellen uit 25 verschillende muizenstammen onderzocht: stamcellen, voorlopercellen, rode bloedcellen en witte bloedcellen. In elk van deze celtypen werd eerst gekeken welke van de 30.000 genen tot expressie kwamen. Veel van deze genen kwamen niet alleen verschillend tot expressie tussen de celtypen, maar ook tussen de muizenstammen. Door in elk van de celtypen variaties in het DNA van de 25 verschillende muizenstammen te vergelijken met variaties in genactiviteit, konden stukjes DNA gevonden worden die een effect hadden op de activiteit van genen. Deze stukjes DNA bevatten dus waarschijnlijk een soort regelaars van genactiviteit. Sommige stukjes DNA bleken in alle vier celtypen hetzelfde effect te hebben op de activiteit van genen. Er waren echter veel meer stukjes DNA die vooral in één, twee of drie van de vier celtypen een effect hadden op de genactiviteit.

De 25 verschillende muizenstammen zijn vele jaren geleden gemaakt door twee duidelijk verschillende muizenstammen met elkaar te kruisen. Deze muizenstammen verschilden bijvoorbeeld in vachtkleur, maar ook in gemiddelde levensduur en meer specifiek in het aantal bloedvormende stamcellen in het beenmerg. Door het nageslacht van deze muizen herhaaldelijk te kruisen werden uiteindelijk 25 verschillende muizenstammen gecreëerd. Het DNA van elk van deze stammen bestaat uit een unieke mix van het DNA van de ouderstammen. Met behulp van zogenaamde moleculaire merkers kon worden vastgesteld welk deel van het DNA van deze stammen afkomstig was van elk van de twee ouderstammen.

Het is bekend dat variaties in het DNA een rol kunnen spelen bij het ontstaan van ziektes zoals leukemie. Het onderzoek in dit proefschrift laat zien dat variaties in het DNA niet altijd hetzelfde effect hebben op de activiteit van genen, maar dat dit effect sterk afhankelijk is van het celtype. Dit betekent dat in toekomstige genetische studies zoveel mogelijk verschillende celtypen bestudeerd moeten worden. Dat is de enige manier om goed te kunnen onderzoeken hoe variaties in het DNA kunnen leiden tot variaties in genactiviteit of zelfs tot de ontwikkeling van ziektes.

Bij de activiteit van genen kan het ook zo zijn dat meerdere (samenwerkende) regelaars een rol spelen. Daarom is er in hoofdstuk 4 ook gezocht naar genen waarvan de activiteit beïnvloed wordt door meerdere regelaars tegelijkertijd. Door variaties in het DNA van de 25 verschillende muizenstammen op een andere manier te vergelijken met variaties in genactiviteit, konden enkele (groepen) genen gevonden worden die door meer dan één regelaar beïnvloed worden.

De stamcellen die bestudeerd zijn in deze studie kunnen zich ontwikkelen tot voorlopercellen die zich op hun beurt weer kunnen ontwikkelen tot rode bloedcellen of witte bloedcellen. Door voor elk van de vier celtypen de genen



– en hun mogelijke regelaars – die in dat celtype tot expressie kwamen in kaart te brengen, konden netwerken gecreëerd worden die betrokken zijn bij het bloedvormingsproces. Zo is er in **hoofdstuk 5** een groep genen gevonden die heel actief is in stamcellen, veel minder actief is in voorlopercellen en nauwelijks meer actief is in rode en witte bloedcellen. Dit doet vermoeden dat de activiteit van deze genen verantwoordelijk zou kunnen zijn voor de unieke eigenschappen van stamcellen. In deze groep genen werd één gen gevonden dat niet alleen heel actief was in stamcellen, maar ook nog eens een mogelijke regelaar was van de activiteit van een aantal andere genen in deze groep. Dit gen was *Zfp521*.

Opmerkelijk was dat er ook al ander bewijs was dat *Zfp521* een belangrijke rol in stamcellen zou kunnen spelen. Naast dat variaties in het DNA van de 25 verschillende muizenstammen vergeleken kunnen worden met de activiteit van genen, kunnen ze ook vergeleken worden met variaties in specifieke eigenschappen van deze muizen. In het verleden was op deze manier al een stukje DNA – met daarin een mogelijke regelaar – gevonden dat verantwoordelijk zou kunnen zijn voor variaties in het aantal stamcellen in het beenmerg van deze muizen. *Zfp521* was één van die mogelijke regelaars. Op dit moment wordt verder onderzocht of *Zfp521* inderdaad een belangrijke rol speelt in stamcellen.

MicroRNAs zijn minuscule afgeleide deeltjes van RNA die in staat zijn om aan RNA te binden en dit te inactiveren, waardoor dit RNA niet meer omgezet kan worden in eiwit. Specifieke microRNAs herkennen specifieke RNA moleculen. In **hoofdstuk 6** werd onderzocht of verschillen in microRNA expressie tussen muizenstammen verantwoordelijk zouden kunnen zijn voor de verschillen in genactiviteit en de verschillen in het aantal bloedvormende stamcellen in het beenmerg van deze muizen. Om dit te kunnen bepalen werd eerst de expressie van alle microRNAs gemeten in de eerder genoemde vier celtypen uit de verschillende muizenstammen. Er werden veel microRNAs gevonden die verschillend tot expressie kwamen tussen celtypen en enkele die verschillend tot expressie kwamen tussen muizenstammen. In deze studie werd een heel interessant microRNA cluster gevonden, bestaande uit drie microRNAs: 99b, let-7e en 125a. Dit cluster was heel actief in stamcellen, minder actief in voorlopercellen en nauwelijks meer actief in gespecialiseerde bloedcellen. Ook was het cluster het meest actief in de muizenstam met de meeste bloedvormende stamcellen in het beenmerg. Deze twee bevindingen impliceerden dat dit cluster een belangrijke rol zou kunnen spelen in stamcellen.

De functie van dit microRNA cluster werd onderzocht door het tot overexpressie te brengen in beenmergcellen en deze cellen vervolgens op verschillende manieren te testen. In celkweek experimenten werd gevonden dat overexpressie van het microRNA cluster cellen vasthoudt in een soort stamcel-staat. Verder werd gevonden dat de bloedvorming in muizen die getransplanteerd werden met deze cellen ontspoorde: sommige muizen ontwikkelden zelfs leukemie. Ook zijn in deze studie een aantal RNA moleculen gevonden waaraan de microRNAs

uit het cluster zouden kunnen binden, wat het omzetten van deze RNA moleculen in eiwit zou blokkeren. Mogelijk is het microRNA cluster via deze RNA moleculen in staat om de keuzes van stamcellen te beïnvloeden.

Geïntroduceerde genetische variatie

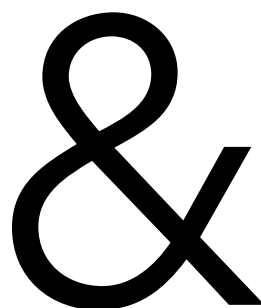
Naast dat de bloedvorming bestudeerd kan worden door gebruik te maken van natuurlijke genetische variatie, kan genetische variatie ook geïntroduceerd worden in het systeem. In **hoofdstuk 7** is genetische variatie in bloedvormende stamcellen van muizen geïntroduceerd door random DNA volgordes of "barcodes" te laten integreren in het DNA van deze cellen. Dit resulteert erin dat elke cel voorzien wordt van een uniek label dat overgegeven wordt van moeder- op dochtercel. De DNA volgorde van het label kan gelezen worden met behulp van een zogenaamde sequencing techniek. Op deze manier kan het lot van bloedvormende stamcellen en hun nakomelingen gevolgd worden na transplantatie in muizen. Deze techniek is zeer veelbelovend voor het monitoren van bloedvorming in toekomstige klinische genterapie studies in de mens.

Een gouden combinatie voor de toekomst?

Een interessante toekomstige studie zou zijn om de bloedvorming te bestuderen door tegelijkertijd gebruik te maken van natuurlijke en geïntroduceerde genetische variatie. Zo zouden stamcellen afkomstig uit verschillende muizenstammen voorzien kunnen worden van een barcode. Deze barcode maakt dat het lot van deze stamcellen na transplantatie gevolgd kan worden. Op deze manier zou de invloed van genetische variatie op specifieke stamceleigenschappen op een heel dynamische manier bestudeerd kunnen worden. Nog een stap verder zou zijn om naast een barcode ook een oncogen in te bouwen in deze stamcellen. Het oncogen maakt dat de bloedvorming op den duur ontspoot. Op deze manier zou ook de invloed van genetische variatie op het ontstaan en de ontwikkeling van leukemie bestudeerd kunnen worden.



DANKWOORD



Eindelijk is het dan zover: mijn proefschrift is een feit! Het voelt alsof ik in een extreme achtbaanrit heb gezeten; een zinderende attractie vol hoogte-, maar soms ook dieptepunten, die elkaar continu afwisselden. Nu deze rit langzaam aan ten einde komt, is het hoog tijd voor een dankwoord. Het schrijven van een dankwoord is echter zo gemakkelijk nog niet, omdat het lastig is om gevoelens van dankbaarheid in woorden uit te drukken. Toch wil ik dat graag proberen.

Gerald, bedankt dat je het vertrouwen in me stelde om me als AiO aan te nemen en me te introduceren in de wondere wereld van de stamcellen, hematologie en genetica. Dankzij jou heb ik in de afgelopen jaren vele aspecten die komen kijken bij het verrichten van wetenschappelijk onderzoek mogen ervaren. Ik prijs mij gelukkig dat ik me met mijn onderzoek op vele niveaus heb mogen begeven: van DNA tot (micro)RNA tot eiwit tot cel tot organisme. Ik heb me dan ook een velerlei aan technieken eigen mogen maken, en zelfs mogen snuffelen aan de bioinformatica. Ik dank je dat je me deze vrijheid hebt gegeven. Bedankt ook dat je me de kans gegeven hebt om mijn werk te presenteren op tal van (inter)nationale congressen. Op deze congressen heb ik altijd vele inspirerende mensen mogen ontmoeten. Je deur stond altijd open om van gedachten te wisselen over het onderzoek. Liep ik bijvoorbeeld even binnen met de vraag of ik beter A of B kon doen, was het antwoord vaak: A én B en misschien ook nog C, D en E. Er waren immers nog zoveel vragen onbeantwoord. Je enthousiasme voor de wetenschap werkte altijd zeer aanstekelijk. Je deur stond ook altijd open om van gedachten te wisselen over persoonlijke zaken. Dat heb ik altijd erg gewaardeerd. **Gerald**, bedankt voor een onvergetelijke tijd!

Leonid, thank you for being my daily supervisor from the start and for the countless scientific discussions we had. Your scientific and technical knowledge is unsurpassed, and combined with your 'thinking outside the box' mentality this has led to the development of new ideas, concepts and techniques. These were invaluable for the research described in this thesis. I also thank you for being critical towards my work; it helped me strive for improvement.

Graag wil ik de leden van de beoordelingscommissie, **Prof. dr. C. Wijmenga**, **Prof. dr. J.N.J. Philipsen** en **Prof. dr. I. Röder**, bedanken voor het lezen en beoordelen van mijn proefschrift.

In het bijzonder wil ik mijn paranimfen **Ellen** en **Lisette** bedanken. Fijn dat jullie naast me willen staan op deze speciale dag. Ellen, een groot deel van het werk dat beschreven staat in dit boekje komt mede uit jouw handen. Bedankt voor de altijd fijne en gezellige samenwerking. Lisette, soms leer je via je opleiding/werk mensen kennen waar je zo'n klik mee hebt dat het vrienden worden en blijven. Sinds de dag dat ik op het MDL lab kwam, klikte het tussen ons. Ik hoop nog vaak samen (met Arjan en Pim natuurlijk) af te spreken!

Veel dank ben ik verschuldigd aan mijn stamcelbiologie collega's met wie ik onderzoeks-lief-en-leed heb gedeeld. **Brad**, I couldn't have wished for a better roommate. I feel lucky to have had you as a coach, mentor, but most importantly as a friend during my PhD studies. Thank you for your everlasting enthusiasm

and optimism and for your kind words when I needed them. I wish you the best of luck in your career, wherever it may take you. Without any doubt life will have great things in store for you. **Sandra O**, wereldrecordhoudster CAFC scoren! Volgens mij ken ik geen hardere werkster dan jij! Bedankt voor al je hulp en onze fijne gesprekken. **Mathilde**, bedankt voor de honderden - wat zeg ik - misschien wel duizenden, prepjes! Als je nog geen wereldrecordhoudster bent, ben je in ieder geval hard op weg om dat te worden. Bedankt voor alle keren dat we samen gelachen hebben. **Jaring**, bedankt voor je hulp bij de vele FACS kleuringen en analyses en voor je FACS-for-dummies protocollen toen je de wijde wereld verkoos boven het stamcellab. **Bertien**, jij bent één van de mensen die het stamcellab draaiende houdt. Bedankt voor je hulp bij grote experimenten en bij alle bureaucratische muizen-rompslomp. **Bert**, je dacht toch niet dat ik je vergeten was? Ik heb het altijd ontzettend fijn gevonden om met je samen te werken, en heb onze leuke gesprekken en je humor dan ook erg gemist nadat je van je pensioen bent gaan genieten. **Martha**, gelukkig was jij er om het stokje van Bert over te nemen. Bedankt voor je goede zorgen voor de dieren en voor de gezellige muizenurtjes. **Erik**, waar ik soms beren op de weg zag bij het analyseren van grote datasets, zag jij daarin juist een uitdaging. Dat is wat jou waarschijnlijk een bioinformaticus maakt, en mij (slechts) een bioloog. Veel plezier en succes bij het verder analyseren en integreren van alle datasets! **Marta, Evgenia, Karin and Visnja**, thank you for the enjoyable time in the lab and at scientific meetings (at locations ranging from Schiermonnikoog to Australia). Good luck finishing your PhDs (I am looking forward to reading your booklets) and in the next phase of your lives! **Marta**; I couldn't wish for a better person to continue with the microRNA project. It feels good to know that I left everything in good hands. **Sandra R**, we started our PhD studies around the same time, visited a lot of (inter)national scientific meetings together and will now also defend our thesis around the same time. Thank you for your support and friendship throughout these years. Your ability to successfully balance work and family life is truly inspirational. I wish you and your beautiful family a bright and happy future, and I hope we will stay in touch. **Ronald**, bedankt voor je input in mijn project, maar ook voor de nodige vrolijkheid (o.a. het zingen) op de werkvloer. **Hein en Evert-Jan**, jammer dat ik maar zo kort met jullie heb mogen werken. Dan zijn daar nog **Piet, Vincent, Kyrjon, Leonie, Joyce, Anita, Isabelle, Esmee, Olya, Marcel en Tobias**, bedankt voor jullie input en gezelligheid! Niet te vergeten; **Annet, Gerry** (en later ook **Yvon**) en **Jan**, jullie rol was misschien vaak op de achtergrond, maar o zo belangrijk. Bedankt voor jullie ondersteuning. En last but not least, natuurlijk dank aan de studenten die ik (mede) heb mogen begeleiden: **Emiel, Joke, Rinse en Bjorg**. Het ga jullie allemaal goed!

Ook buiten de stamcelbiologie heb ik veel hulp gehad. Ik wil graag mijn collega's van de **Straling & Stress Celbiologie** bedanken voor de prettige tijden op het lab en daarbuiten. Dank ook aan mijn collega's van de **Hematologie** (in het bijzonder wil ik hier **Edo, JJ, Gerwin** en **Bart-Jan** noemen) en de **Kinderoncologie**



voor jullie input en de leerzame discussies tijdens de maandagochtend seminars. Dan zijn daar nog de medewerkers van het CDP, bedankt voor jullie ondersteuning en de goede zorgen voor de dieren! Daarnaast wil ik graag **Henk, Geert, Harold** en **Roelof-Jan** bedanken voor alle celsorteringen. Wat heb ik vele uren met jullie doorgebracht! Toch was er altijd wel wat om over te praten. Zonder jullie hadden veel van de experimenten beschreven in dit proefschrift niet uitgevoerd kunnen worden en had dit proefschrift er heel anders uitgezien. **Bruno, Yang, Rainer** and **Ritsert**, thank you for introducing me into the world of bioinformatics. Although sometimes it may have felt that we were speaking different languages – biology versus bioinformatics – I think this type of interdisciplinary research is the science of the future! **Jeroen**, ik heb de samenwerking met jou als heel prettig ervaren; bedankt voor je frisse kijk op het onderzoek! **Joost** en **Anke**, microRNAs waren eerst een gesloten boek voor mij. Bedankt dat jullie dat voor mij hebben geopend. **Bart** en **Susanne**, we hebben dan wel niet echt samengewerkt, maar wel een geweldige ervaring gedeeld in de jungle van Australië! A special thanks goes to the Leipzig/Dresden group: **Ingo** (see above), **Ingmar, Tilo** and **Nico**, we visited many scientific meetings together and somehow we always ended up having dinner or drinks together. Thank you for the many great discussions we had about science and the world outside of science. I wish you all the best for everything that is to come!

Kort wil ik nog het woord richten tot de mensen die de Topmaster MPDI tot een groot succes gemaakt hebben. Bedankt voor jullie bijdrage aan mijn wetenschappelijke vorming. Het waren twee ontzettend leuke jaren, vol met uitdagingen. **Anke**, dankzij jou was alles altijd goed georganiseerd. **Han**, bedankt voor de leuke en leerzame tijd op het MDL lab! **Harrie**, ook tijdens mijn promotieonderzoek bleef je geïnteresseerd in hoe het me verging. Ik heb onze gesprekken altijd als heel fijn ervaren. **Elise, Willemien** en **Marieke**, bedankt voor de gezelligheid tijdens ons Topmaster avontuur!

Naast oud-collega's in het UMCG in Groningen, zijn er ook nieuwe collega's in de **Isala Klinieken** in Zwolle. Medewerk(st)ers van het **Klinisch Chemisch Laboratorium**, mijn start bij jullie liep helaas heel anders dan gedacht. Ik wil jullie graag bedanken voor jullie begrip en voor de ruimte die jullie me de afgelopen periode gegeven hebben. In het bijzonder wil ik hier **Bert, Robbert, Winy, Ilse, Niels** en **Michel** noemen, bedankt dat jullie me opgenomen hebben in jullie groep. Ik hoop dat ik in Zwolle net zo'n leuke en leerzame tijd mag hebben als in Groningen. **Erna**, harde werkster! Heel veel succes (en natuurlijk plezier) met het verdedigen van je (mooie!) proefschrift!

Hoewel de eerder genoemde personen vooral een directe invloed hebben gehad op de inhoud van dit proefschrift, was ook de indirecte steun van vrienden en familie onontbeerlijk voor het uiteindelijke resultaat. **Annemiek** en **Gerjan, Lisette** (zie boven) en **Arjan, Karin, Marianne** en **Rolf, Peter** en **Chantal, Minke, Jeroen, Jessie, Anneke K, Martine, Mieke, Anneke Z** en **Aukje**, ook al wonen jullie

allemaal een eindje weg en zie ik jullie lang niet zo vaak als dat ik zou willen: jullie vriendschap is me erg dierbaar. Ik wil jullie bedanken voor alle mooie momenten die we tot dusver al hebben gehad en alvast voor alle mooie momenten die nog komen gaan! **Familie Hekman**, bedankt voor jullie steun en interesse de afgelopen jaren. Dat er nog maar vele eier-eet-competities en barbecues mogen volgen! Lieve schoonouders, **Harm** en **Truus**, geweldig hoe jullie ons helpen. Niets is teveel, altijd staan jullie voor ons klaar.

Jan Gerard, je bent mijn enige broer(tje) en daarom natuurlijk ook meteen de liefste. Hoewel je het altijd ontzettend druk hebt, heb je (of maak je) altijd tijd om anderen te helpen. Zo ook mij. Bedankt daarvoor! **Nienke**, ik ken je nog niet zo lang, maar wilde je hier toch noemen. Ik hoop je snel beter te leren kennen. Ik wens jullie alle liefde en geluk van de wereld!

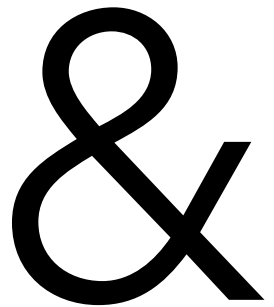
Lieve **Robert**, dankzij jou is dit proefschrift nu een feit. Als geen ander heb jij van dichtbij gezien hoe dit proefschrift vorm kreeg. Bedankt voor je bijdrage aan dit proefschrift, maar veel meer nog voor alle andere (niet-proefschrift-gerelateerde) grote en kleine dingen. Veel is er het afgelopen jaar gebeurd en jij bleef mijn rots in de branding. Hiervoor ben ik je onbeschrijflijk dankbaar. Bedankt voor alles wat was, is en nog komen zal.

Lieve **Papa** en **Mama**, jullie onvoorwaardelijke liefde, steun, vertrouwen en betrokkenheid vormen de basis voor dit alles. Daarom draag ik mijn boekje aan jullie op. Jullie hebben me altijd gestimuleerd om te worden wat ik wilde worden, om te doen wat ik wilde doen en om te zijn wie ik ben. Lieve Papa, het valt onmogelijk in woorden uit te drukken hoeveel je voor mij betekend hebt en nog steeds betekent. Jij bent in alles mijn inspiratie geweest en dat zal je ook altijd blijven. Wat hebben we het vaak gehad over mijn experimenten, artikelen, proefschrift, stellingen en ook mijn promotie. Wie had toen kunnen voorspellen dat jij deze promotie niet meer mee zou maken. Ik zal jou en je trotse glimlach ontzettend missen op deze speciale dag. Je hebt nog net meegemaakt dat ik in Zwolle begonnen ben als klinisch chemicus in opleiding. Ook dit nieuwe avontuur had ik graag met jou willen aangaan. Juist met jou. Het zou een prachtig avontuur geworden zijn. Lieve Papa, ik ben ongelooflijk dankbaar en trots dat ik jouw dochter heb mogen zijn. In mijn hart leef je nog volop. Lieve Mama, bedankt dat je mij (en JG) altijd zo'n fijn thuis gegeven hebt. Je stond (en staat) altijd klaar met een advies of gewoon een luisterend oor. Ik had me geen betere moeder kunnen wensen! Liefste Mama, samen staan we sterk en slaan we ons erdoor.

Alice Gerrits
Zwolle, juli 2011



BIOGRAFIE
BIOGRAPHY
LIST OF PUBLICATIONS
ATTENDED MEETINGS
FUNDING AND AWARDS





BIOGRAFIE

Alice Gerrits werd geboren op 29 juli 1981 te Hardenberg. Na het behalen van haar VWO diploma aan het Vechtdal College te Hardenberg in 1999, begon zij met de opleiding 'Chemie, Biologie en Laboratoriumonderzoek' aan de Hogeschool Drenthe te Emmen. Na het behalen van haar propedeuse vervolgde zij in 2000 haar studie aan de Rijksuniversiteit Groningen met de opleiding 'Medische Biologie'. In 2003 rondde zij deze opleiding *cum laude* af met een Bachelor diploma, waarna zij geselecteerd werd voor de Topmaster opleiding 'Medical and Pharmaceutical Drug Innovation' van de Rijksuniversiteit Groningen, waar zij in een intensief 2-jaars programma voorbereid werd op een solide start in het onderzoek. Tijdens haar Master opleiding heeft ze twee onderzoeksprojecten afgerond bij de afdeling Maag-, Darm en Leverziekten en de afdeling Stamcelbiologie, beide in het Universitair Medisch Centrum Groningen. In het kader van deze opleiding heeft ze onder begeleiding van Prof. dr. Gerald de Haan een voorstel voor een AiO-project geschreven, welke werd gehonoreerd door de onderzoeksschool GUIDE. In september 2005 rondde ze de Topmaster opleiding *cum laude* af, waarna zij begon als promovenda op het bovengenoemde project bij de afdeling Stamcelbiologie in het Universitair Medisch Centrum Groningen. In haar onderzoek maakte ze gebruik van natuurlijke en geïntroduceerde genetische variatie om het bloedvormend systeem te bestuderen. Dit onderzoek vond plaats in de context van twee Europese Unie projecten: 1) EUrythron, met de focus op de aanmaak van rode bloedcellen, en 2) EuroSyStem, met de focus op stamcellen en systeem biologie. De resultaten van haar promotieonderzoek staan beschreven in dit proefschrift. Sinds november 2010 werkt Alice als klinisch chemicus in opleiding in de Isala klinieken te Zwolle.



BIOGRAPHY

Alice Gerrits was born on July 29, 1981 in Hardenberg, The Netherlands. After receiving her pre-university degree from the Vechtdal College in Hardenberg in 1999, she spent one year in the 'Chemistry, Biology and Medical Laboratory Research' program at the University of Applied Sciences in Emmen. Starting in 2000 she studied 'Medical Biology' at the University of Groningen, obtaining her Bachelor's degree *cum laude* in 2003. Thereafter, she was selected to participate in the Topmaster program 'Medical and Pharmaceutical Drug Innovation' of the University of Groningen, in which she – in an intensive two-year program – was prepared for a solid start in research. During her Masters studies she completed research projects at the department of Gastroenterology/Hepatology and the department of Stem Cell Biology, both in the University Medical Center Groningen, and graduated *cum laude* in 2005. In the final stage of her Topmaster program she wrote a PhD project proposal under supervision of Prof. dr. Gerald de Haan, which was funded by the Graduate School GUIDE. As a result, in September 2005 she began her PhD studies at the department of Stem Cell Biology in the University Medical Center Groningen. In her research project she made use of natural and induced genetic variation to study the process of blood cell formation. This project took place in the context of two European Union projects: 1) EUrythron, with the focus on red blood cell formation, and 2) EuroSyStem, with the focus on stem cells and systems biology. The results of her PhD research are presented in this dissertation. In November 2010 Alice entered a clinical chemistry training program in the Isala Clinics in Zwolle.

&

LIST OF PUBLICATIONS

1. De Haan G, Gerrits A, Bystrykh L. Modern genome-wide genetic approaches to reveal intrinsic properties of stem cells. *Curr Opin Hematol*. 2006 Jul;13(4):249-53. Review.
2. De Haan G, Gerrits A. Epigenetic control of hematopoietic stem cell aging - the case of Ezh2. *Ann N Y Acad Sci*. 2007 Jun;1106:233-9. Review.
3. Gerrits A, Dykstra B, Otten M, Bystrykh L, De Haan G. Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny. *Immunogenetics*. 2008 Aug;60(8):411-22. Review.
4. Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, Gerrits A, Bystrykh LV, De Haan G, Su AI, Jansen RC. Genetical genomics: spotlight on QTL hotspots. *PLoS Genet*. 2008 Oct;4(10):e1000232.
5. Gerrits A*, Li Y*, Tesson BM*, Bystrykh LV, Weersing E, Ausema A, Dontje B, Wang X, Breitling R, Jansen RC, De Haan G. Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genet*. 2009 Oct;5(10):e1000692.
* *Equal contribution*
** *Highlight in Nature Rev Genet*. 2009 10 (12):819
6. Gerrits A, Dykstra B, Kalmykova OJ, Klauke K, Verovskaya E, Broekhuis MJC, De Haan G, Bystrykh LV. Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood*. 2010 Apr 1;115(13):2610-8.
** *Appeared on the cover of Blood*
7. De Ridder J, Gerrits A, Bot J, De Haan G, Reinders M, Wessels L. Inferring combinatorial association logic networks in multimodal genome-wide screens. *Bioinformatics*. 2010 Jun 15;26(12):i149-57.
8. Gerrits A, Walasek MA, Olthof S, Weersing E, Ritsema M, Zwart E, Bystrykh LV, De Haan G. Genetic screen identifies microRNA cluster 99b/let-7e/125a as a regulator of primitive hematopoietic cells. *Invited for resubmission to Blood*.
9. Gerrits A, Li Y, Tesson BM, Breitling R, Göttgens B, Jansen RC, Bystrykh LV, De Haan G. Genetic screen identifies *Zfp521* as a candidate regulator of hematopoietic stem cell pool size. *In preparation*.



ATTENDED MEETINGSⁱⁱ

1. Annual International Society for Hematology and Stem Cells (ISEH) meetings
2007, Hamburg, Germany (O); 2010, Melbourne, Australia (O)
2. Annual International Society for Stem Cell Research (ISSCR) meetings
2006, Toronto, Canada (P); 2007, Cairns, Australia (O); 2008, Philadelphia,
US (O)
3. Annual Stem Cells in Development and Disease (SCDD) meetings
2007, Amsterdam; 2008, Amsterdam (P); 2009, Rotterdam (O)
4. Annual Dutch Hematology meetings
2005, Lunteren; 2007, Arnhem (O); 2009, Arnhem (O)
5. EUrythron meetings
 - Annual consortium meetings
2006, Rotterdam (O); 2007, Paris, France (O); 2008, Groningen (O);
2009, Lisbon, Portugal (O)
 - Workshop molecular control of erythropoiesis, 2005, Rome, Italy (O)
 - Career event in molecular cell biology and medicine, 2008, Porto,
Portugal
6. EuroSyStem meetings
 - Annual consortium meetings
2009, Cambridge, UK (P); 2010, Schiermonnikoog (O)
 - Workshop computational approaches in stem cell biology, 2009,
Leipzig, Germany (O)
7. Other meetings
 - Dutch Chromatin Meeting, 2006, Rotterdam (O)
 - Complex Trait Consortium meeting, 2007, Braunschweig, Germany (P)
 - Royal Netherlands Academy for Arts and Sciences (KNAW) meeting:
role of DNA polymorphisms in complex traits and diseases, 2008, Amsterdam
 - European summer school: stem cells and regenerative medicine, 2008,
Hydra, Greece (P)
 - European Molecular Biology Organization (EMBO) meeting: stem cells,
systems and synthetic biology, 2009, Cambridge, UK (P)



ⁱⁱ(O) = Oral presentation, (P) = Poster presentation

FUNDING AND AWARDS

1. 2005 – Graduate School GUIDE funding for PhD project
2. 2006 – ISSCR Junior Investigator poster award
3. 2006 – New York Stem Cell Foundation travel award
4. 2007 – ISEH travel award
5. 2007 – ISEH Young Investigator award/ Dirk van Bekkum award
6. 2007 – ISSCR travel award
7. 2008 – ISSCR travel award
8. 2010 – ISEH travel award



Now this is not the end. It is not even the beginning of the end.
But it is, perhaps, the end of the beginning.

Winston Churchill

