GENE EXPRESSION STUDIES

FROM BASIC RESEARCH TO THE CLINIC

Juha Karjalainen

Cover painting: Helena Vatanen
Cover design, Layout & Printing: Lovebird design
www.lovebird-design.com

ISBN (printed version): 978-94-034-0317-5

university of groningen

Gene expression studies from basic research to the clinic

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 15 January 2018 at 16.15 hours

by

Juha Matti Karjalainen

born on 28 March 1983
in Mikkeli, Finland

Supervisors
Prof. T.N. Wijmenga
Prof. L.H. Franke

Assessment Committee
Prof. H. Snieder
Prof. E.C. Wit
Prof. M.J.T. Reinders

Vanhemmilleni

TABLE OF CONTENTS

1 Preface and outline 9

2 Gene expression analysis identifies global gene dosage sensitivity in cancer 17

3 Human disease-associated genetic variation impacts large intergenic non-coding RNA expression 33

4 Biological interpretation of genome-wide association studies using predicted gene functions 45

5 Discussion and future perspectives 57

6 Summaries 71

7 Abbreviations 79

8 Acknowledgements 83

9 List of publications 89

1

PREFACE AND OUTLINE

Eläin elää, ihminen ihmettelee

Sielun Veljet — Säkenöivä voima

1.1 PREFACE

Genetics as a field of scientific study has existed for a century. However, theories of heredity have been laid out for millennia. The Greek philosophers and mystics suggested how features, or phenotypes, were passed on through human generations. For instance, Pythagoras (c. 550 BC) believed, with the lack of anatomical evidence, that the male semen circulated through the body and mystically collected information from its parts. The semen would contain instructions for the creation of a human being, whereas the female provided for its growth — that the nature came from the male and the nurture from the female.[1,2]

Long after Pythagoras' time, the Czech friar Johann Mendel proposed a remarkable theory of discrete units of heredity in his 1866 paper entitled "Versuche über Pflanzen-Hybriden" (Experiments on Plant Hybridization).[3] Mendel cross-hybridised pea plants over the course of eight years and observed interesting phenomena over the generations. When he crossed short plants with tall plants, only tall plants would result. But when he crossed these second-generation plants with each other, some of the resulting plants would be short. Repeating these experiments and observing several characteristics of the plants, Mendel noticed that the characteristics of later generations would appear in beautiful mathematical ratios and independently of each other: the height of the plant was inherited regardless of the colour of the seed, for example. While Mendel was not one for schoolbooks,[4] he postulated the laws of segregation, independent assortment, and dominance, which are the foundations of today's schoolbook genetics.

Several years before Mendel's pea experiments, in 1831–1836, the English naturalist, geologist and biologist, Charles Darwin, had sailed to the islands and coasts of South America where he collected fossils and animal corpses to be shipped back to Europe. He observed that species on different islands were slightly different from each other: "Each variety is constant on its own island".[5,6] His observations and thinking led him to postulate his theory of natural selection and "survival of the fittest", a phrase borrowed from Herbert Spencer, an English philosopher and scientist.[7]

If Darwin was aware of Mendel's work, he didn't connect it to his own. When working on his theory in the decades after the voyage, Darwin struggled with how inheritance would fit into his theory of evolution: Why didn't "freaks of nature", such as short beaks of birds, disappear from populations over time? He had assumed that the characteristics of individuals blended together over generations, and that characteristics acquired by an individual during its lifetime were inherited, much as Pythagoras had thought 2000 years earlier.

Among others, the Dutch botanist Hugo de Vries first rediscovered Mendel's findings and then his work at the turn of the twentieth century, publishing his own work in 1901.[8] Based on his experiments with primroses, de Vries realised that mutations such as relatively big leaves were passed on to the next generations discretely — not in a blending fashion as Darwin had assumed. Mendel's findings of generation-to-generation reproduction and Darwin's theory of long-term evolution converged into a new field of study, which the English biologist William Bateson first called "genetics" in 1905.[9]

While the basic principles of heredity had been discovered, its molecular basis was still unclear in the early twentieth century. Microscopic knowledge of the late nineteenth century had led scientists to hypothesise that the genetic material resides in the nuclei of cells. The existence of chromosomes had been shown, leading to the discovery of the cell reproduction mechanisms of mitosis and meiosis. It was accepted that the concept of a gene had a physical and biological form, but it was unclear what this form was. The American geneticist Thomas Morgan worked with fruit flies and concluded in his 1911 paper, entitled "Random segregation versus coupling in Mendelian inheritance", that some characteristics are linked to each other in contrast to Mendel's law of independent assortment. He also found that some characteristics are inherited on a sex chromosome, and that genes are carried on chromosomes in a sequential manner.[10]

The existence of deoxyribonucleic acid (DNA) in cells was recognized a few years after Mendel's pea experiments. However, its purpose remained unclear for nearly a century. During the first decades of the twentieth century, biochemists found that DNA consisted of four different nitrogen bases strung together. In 1944, the Canadian-American physician Oswald Avery discovered from his experiments that genes may reside within the DNA molecule.[11] Scientists then started racing to determine the physical structure of DNA. In England, heavily influenced by the chemist Rosalind Franklin's work on making X-ray photographs of DNA, the biologist James Watson and biophysicist Francis Crick suggested its double-helix structure, in which the nitrogen bases are unambiguously paired with each other.[12,13] "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material", was Watson and Crick's exciting conclusion to their seminal paper in 1953.

In subsequent decades, research shifted from understanding the form of the DNA to understanding its function. The Indonesian-born American cytogeneticist Joe Hin Tjio and Swedish botanist Albert Levan reported the correct number of chromosomes in humans (46) in 1955, two years after the discovery of the DNA structure.[14] The copying mechanism, or replication, of the DNA molecule was affirmed by the American molecular biologists Matthew Meselson and Franklin Stahl in 1958.[15] By 1965, the basic mechanism of transcription of DNA to ribonucleic acid (RNA) was understood.

Following the concrete discoveries of the fundamental molecular processes of DNA replication, RNA transcription, and protein translation, during the next decades the functions of genes were studied intensely. These studies ranged from the development of organisms to what is today the focus of human genetics: human diseases. In 1972, biologists Herbert Boyer and Stanley Cohen introduced viral DNA into a bacterial cell, marking a starting point for recombinant DNA technology — combining DNA from different organisms — and gene cloning.[16,17]

In 1977, the English biochemist Frederick Sanger invented a gene sequencing technique to reveal the full 5386-nucleotide long DNA sequence of the ΦX174 virus, preparing the path for unravelling the genetic code of other organisms.[18] Meanwhile, in 1980, David Botstein and his colleagues proposed a method to construct a genetic linkage map of the human genome.[19] In 1983, the genetic locus linked to Huntington's disease was found using this method, marking the first discovery of a single disease locus.[20] With the technologies available at the time, it took another ten years of gene mapping to find the culprit, a large protein-coding gene first named "interesting transcript 15", and later huntingtin.[21]

After Sanger's work with the ΦX174 virus, scientists were eager to discover the DNA sequences of more complex organisms. In 1998, the genome of the worm *Caenorhabditis elegans*, better known as *C. elegans*, had been sequenced in Hinxton, near Cambridge, UK.[22] *C. elegans* thus became the first multicellular organism whose genome sequence was revealed. Then, in 2000, a consortium of scientists led by the American biotechnologist Craig Venter, and his company Celera Genomics, published the draft sequence of the fruit fly, a model organism traditionally used in genetics.[23]

To determine the full human DNA sequence, an international collaboration called the Human Genome Project (HGP) had been launched in 1990 in the United States, long before the genomes of *C. elegans* and the fruit fly had been discovered. It was the biggest project in biology ever undertaken. As it progressed, Craig Venter became determined to discover the sequence of the human genome independently of the HGP, using a different sequencing strategy (shotgun sequencing instead of clone based) that Celera had also used in determining the fruit fly genome. The two projects raced forward neck and neck. After a political intervention by the US President Bill Clinton,[24] papers revealing the draft DNA sequences from both projects were published simultaneously in 2001 in rival journals: the HGP's work appeared in Nature and Celera's effort in Science.[25,26] A century after the beginning of genetics, humans had unravelled their own manuscript, which has since been intensively interpreted and led to an increased understanding of biology and human disease.

Meanwhile, by the turn of the millennium, DNA sequencing methods had improved enormously. These "next-generation" or high-throughput techniques process sequences in parallel, speeding up the sequencing procedure tremendously. In addition, the cost of sequencing an individual genome has fallen by several orders of magnitude in less than two decades. The Human Genome Project was eventually finished in 2003, cost nearly $3 billion, and took 13 years to complete. Now, in 2017, a whole human genome can be sequenced for a few hundred dollars in just a few days.[27] This allows for massive studies of hundreds of thousands of individuals.

Next to sequencing, genotyping technologies have been used since 1980 and in the last two decades even more widely due to the availability of genome-wide DNA chips that allow for high-throughput and cheap genotyping.[28] Instead of identifying parts of DNA sequences, genotyping determines genetic variants at specified positions in the genome. Since 2005, a plethora of genome-wide association studies (GWASs) have been performed to identify single nucleotide polymorphisms (SNPs) — variations of a single nucleotide in a specific position of the genome — associated with a disease or phenotype.[29,30] By August 2017, these studies have compared variants between groups of cases and controls, and found nearly 40,000 unique associations between SNPs and various phenotypes.[31]

Despite the success of GWASs in finding genetic variants associated with phenotypes, it is still often unclear how these variants cause a phenotype.[30] For example, today more than a hundred common variants (variants seen in at least 1 % of the population) are known to be associated with schizophrenia.[32] Each of these variants individually explains only a small proportion of susceptibility to the disease — it is believed that this complex disease is caused by hundreds of genetic variants that act together. Back in 1918, statistician Roland Fisher had proposed an infinitesimal model stating how several genes with small effect sizes could contribute to a certain phenotype.[33] But, in most cases, it is still not known what role each individual associated variant or region plays in disease aetiology, or how the variants together cause the disease.

While DNA sequence and genotype information can be seen as the basis of biology, the emergence of life out of DNA is an extremely complex phenomenon. In cells, DNA is transcribed into ribonucleic acid (RNA), which in turn is translated into the proteins that make us up. Francis Crick called this flow of information "the central dogma of molecular biology", although he later regretted the use of the word "dogma".[34] Since Crick's work, we have learned a lot about how genes are regulated — turned "on" or "off" — in various ways, and how they work together in different ways depending on the developmental stage of the organism, the type of the cell, and environmental factors. We have come to appreciate the true complexity of biology, partly because we know so little about the genetics and aetiology of complex diseases. Now we also know that cells contain information that is not encoded in the DNA sequence itself. The field that studies this phenomenon is called "epigenetics". In the last decades, the field of genetics has become multidisciplinary, requiring efforts from biologists, mathematicians, chemists, physicians, informaticians and philosophers. There is far more to study with regards to biology, disease and health, than previously anticipated!

In addition to sequences and genotypes, higher levels of molecular information are now being extensively measured. Gene expression is the process in which genes as stretches of DNA become their final products: RNA or protein. RNA transcription is the first step in this process. As such, measuring and analysing quantities of RNA in cells from various tissues, conditions and diseases can help us in investigating cellular phenomena, as well as disease biology. Traditionally, RNA quantities in cells have been measured using predetermined microarrays. In the course of the last seven years, RNA sequencing (RNA-seq) has become a prevalent method.[35,36] While microarray measurements are inherently limited to the probes present on the array, RNA-seq provides an unbiased, genome-wide view of the abundances of RNA in cells. This allows for investigation of large intergenic non-coding RNAs (lincRNAs) for example, for most of which there are no probes on microarray platforms. Often, and in this thesis, the term "gene expression" refers to RNA transcription.

In this thesis, I study the functions of genes in relation to biological pathways, cell types, tissues and phenotypes, including diseases. When I started my PhD research in February 2011, the National Center for Biotechnology Information's (NCBI) dbSNP database listed 30 million human SNPs. At the same time, the GWAS Catalog of the National Human Genome Research Institute (USA) and the European Bioinformatics Institute (NHGRI-EBI) contained 4,200 disease-associated SNPs. At the time of writing, in August 2017, these numbers had grown to 300 million and 35,000, respectively.[31,37] While our knowledge of genetic variation and its association with various phenotypes has increased, it is still unclear how genetic variation translates to phenotypic variation and disease, as in the above example of schizophrenia. We also do not yet have a comprehensive view of the functions of individual genes and, in many cases, we do not yet know which genes in the disease-associated genetic regions are relevant to the disease biology.

To gain a better understanding of genetics and genomics in these respects, scientists have during the last two decades developed several tools to predict gene function, to find functional enrichment among genes of interest, and to prioritize genes in genetic loci implicated by GWASs.[38,39] PANTHER (Protein Annotation THrough Evolutionary Relationship) is a widely used system for classifying gene and protein function. Initially released in 2003, it is based on phylogenetic data from various organisms combined with functional data. PANTHER determines protein and gene families by sequence homology, as well as subfamilies by shared function.[40,41]

Three widely used tools for analysing lists of genes deemed interesting in biological experiments are GSEA (Gene Set Enrichment Analysis),[42] DAVID (the Database for Annotation, Visualization and Integrated Discovery),[43] and MAGENTA (Meta-Analysis Gene-set Enrichment of variaNT Associations).[44] These tools traditionally rely on known gene functions to find enriched pathways for the user's genes of interest.

In addition to finding functional enrichments for gene lists, researchers have been increasingly eager to determine potentially causal genes based on SNPs found to be associated with a phenotype in a GWAS. GRAIL (Gene Relationships Across Implicated Loci) has been a widespread method for this purpose.[45] It is a text-mining approach that finds similarities in scientific literature among genes in given genetic loci.

Despite the usefulness of the traditional methods for analysing results of genetic studies, they share the shortcoming of relying on previously established biological knowledge. On the other hand, gene function prediction methods have shared limitations of relying on sequence similarity or not being able to consider molecular measurements from a wide variety of conditions, thus lacking means to systematically predict functions genome-wide.

During the last decade, a myriad of measurements of gene expression levels have been made freely available by researchers and technicians around the world. These molecular data has allowed me to jointly study gene expression patterns in various tissues and conditions. I have been able to use these patterns in combination with established knowledge of biological pathways to systematically and accurately predict functions for genes, which has resulted in finding previously unknown gene functions. Additionally, I have used findings from genomewide association studies to successfully prioritise relevant genes and pathways for various phenotypes based on the predicted gene functions. The bioinformatic methods described in this thesis

1.2 OUTLINE OF THE THESIS

My thesis has two main aims: (1) to present a gene expression-based method to predict novel functions for human genes, and (2) to present a follow-up method to suggest which genes and pathways may be relevant to different phenotypes, based on findings from genome-wide association studies.

Chapter 2 describes a method to predict functions for human genes based on gene expression data and previously established gene functions. We reanalysed 77,840 publicly available microarray expression profiles. Using principal component analysis (PCA), we found "transcriptional components" that describe biological phenomena. We used these components together with established pathway information to systematically predict gene functions for all the human genes represented on the microarray platforms. For example, we predicted and validated that the *FEN1* gene is required for homologous recombination-mediated repair. We also used the gene expression data to identify somatic copy number alterations in cancer samples.

Chapter 3 shows that many lincRNAs are under genetic control. We tested the association of SNPs with expression levels of lincRNAs from five different tissues and identified 112 cis-regulated lincRNAs. We noticed that 75 % of SNPs that affected lincRNA expression did not affect the expression of nearby protein-coding genes. Using the method described in chapter 2, we predicted relevant functions for four lincRNAs whose expression was affected by disease-associated SNPs.

Chapter 4 describes a computational method called DEPICT, built on the work in chapter 2, to systematically prioritise relevant genes at disease-associated loci that have been found in genome-wide association studies. The method also suggests pathways that are relevant to the studied phenotype, as well as tissues in which the prioritised genes are highly expressed. DEPICT uses no phenotype-specific hypotheses. We benchmarked it with three different phenotypes: Crohn's disease, human height, and low-density lipoprotein cholesterol. We observed that the method had superior performance compared to existing methods.

In chapter 5, I discuss the findings of my work and place them in a broad perspective. I also consider some future directions for research and some clinical applications of the work described here.

REFERENCES

[1] S. Mukherjee. *The Gene: An Intimate History*. Scribner, 2016.

[2] I. C. Johnston. *...And Still We Evolve — A Handbook for the Early History of Modern Science*. 3 edition, 1999. http://records.viu.ca/~johnstoi/darwin/title.htm [Accessed: 24th Aug 2017].

[3] G. Mendel. Versuche über Pflanzen-Hybriden. V*erhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865, Abhandlungen*, pages 3–47, 1866.

[4] J. Klein and N. Klein. *Solitude of a Humble Genius — Gregor Johann Mendel: Volume 1: Formative Years*. Springer-Verlag Berlin Heidelberg, 2013.

[5] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.

[6] S. Herbert. *Charles Darwin, Geologist*. Cornell University Press, 2005.

[7] H. Spencer. *The Principles of Biology*. Williams and Norgate, 1864.

[8] H. de Vries. *Die mutationstheorie. Versuche und beobachtungen über die entstehung von arten im pflanzenreich*. Leipzig, Veit & comp., 1901.

[9] B. Bateson. *William Bateson, Naturalist: His Essays and Addresses Together with a Short Account of His Life*. Cambridge University Press, 2009.

[10] T. Morgan. Random segregation versus coupling in Mendelian inheritance. *Science*, 34 (873):384, Feb 1911.

[11] O. T. Avery, C. M. MacLeod, and M. McCarty. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types — Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *The Journal of Experimental Medicine*, 79(2):137–158, Feb 1944.

[12] J. D. Watson, A. Gann, and J. Witkowski. *The Annotated and Illustrated Double Helix*. Simon & Schuster, 2012.

[13] J. D. Watson and F. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171:737–738, Apr 1953.

[14] J. H. Tjio and A. Levan. The chromosome number of man. *Hereditas*, 42:1–6, May 1956.

[15] M. Meselson and F. W. Stahl. The Replication of DNA in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 44(7):671–682, Jul 1958.

[16] S. Cohen, A. Chang, H. Boyer, and R. Helling. Construction of biologically functional bacterial plasmids in vitro.

Proceedings of the National Academy of Sciences of the United States of America, 70(11):3240–3244, Nov 1973.

[17] S. Cohen. DNA cloning: A personal view after 40 years. *Proceedings of the National Academy of Sciences of the United States of America*, 110(39):15521–15529, Sep 2013.

[18] F. Sanger et al. Nucleotide sequence of bacteriophage ΦX174 DNA. *Nature*, 265: 687–695, Feb 1977.

[19] D. Botstein et al. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314–331, May 1980.

[20] J. F. Gusella et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306:234–238, Nov 1983.

[21] M. MacDonald et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–983, Mar 1993.

[22] C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, 282(5396):2012–2018, Dec 1998.

[23] M. Adams et al. The genome sequence of Drosophila melanogaster. *Science*, 287(5461): 2185–2195, Mar 2000.

[24] CNN. The race is over. Jun 2000. http://edition.cnn.com/ALLPOLITICS/time/2000/06/26/race.html [Accessed: 29th Aug 2017].

[25] E. S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409: 860–921, Jan 2001.

[26] J. C. Venter et al. The Sequence of the Human Genome. *Science*, 291(5507):1304–1351, Feb 2001.

[27] National Institutes of Health — National Human Genome Research Institute. The Cost of Sequencing a Human Genome. 2016.
https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome
[Accessed: 24th Aug 2017].

[28] R. Bumgarner. DNA microarrays: Types, Applications and their future. *Current Protocols in Molecular Biology*, 101:22.1.1–22.1.11, Jan 2013.

[29] R. J. Klein et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720):385–389, Apr 2005.

[30] P. M. Visscher et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101(1):5–22, Jul 2017.

[31] J. MacArthur et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45 (Database issue):D896–D901, 2017.

[32] Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511:421–427, Jul 2014.

[33] R. Fisher. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, 52:399–433, Oct 1918.

[34] R. Olby. *Francis Crick: Hunter of Life's Secrets*. Cold Spring Harbor Laboratory Press, 2009.

[35] S. Goodwin et al. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, May 2016.

[36] M. Ritchie et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleid Acids Research*, 43(7):e47, Apr 2015.

[37] S. Sherry et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, Jan 2001.

[38] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleid Acids Research*, 37(1):1–13, Jan 2009.

[39] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *American Journal of Human Genetics*, 86(1):6–22, Jan 2010.

[40] P. D. Thomas et al. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, 13(9):2129–2141, Sep 2003.

[41] H. Mi et al. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, 8(8):1551–1566, Aug 2013.

[42] A. Subramanian. Gene set enrichment analysis: A knowledge-based approach for in-terpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, Oct 2005.

[43] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, Jan 2009.

[44] A. V. Segrè et al. Common Inherited Variation in Mitochondrial Genes is not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. *PLOS Genetics*, 6(8), Aug 2012.

[45] S. Raychaudhuri et al. Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. *PLOS Genetics*, 5(6), Jun 2009.

2

GENE EXPRESSION ANALYSIS IDENTIFIES GLOBAL GENE DOSAGE SENSITIVITY IN CANCER

Nature Genetics, 47:115-125, Jan 2015

Rudolf S. N. Fehrmann, Juha M. Karjalainen, Małgorzata Krajewska,
Harm-Jan Westra, David Maloney, Anton Simeonov, Tune H. Pers,
Joel N. Hirschhorn, Ritsert C. Jansen, Erik A. Schultes,
Herman H. H. B. M. van Haagen, Elisabeth G. E. de Vries, Gerard J. te Meerman,
Cisca Wijmenga, Marcel A. T. M. van Vugt & Lude Franke

Supplementary information is available at https://goo.gl/JqpCya

nature genetics

Gene expression analysis identifies global gene dosage sensitivity in cancer

Rudolf S N Fehrmann[1,2,12], Juha M Karjalainen[2,12], Małgorzata Krajewska[1], Harm-Jan Westra[2], David Maloney[3], Anton Simeonov[3], Tune H Pers[4–7], Joel N Hirschhorn[4–6,8], Ritsert C Jansen[9], Erik A Schultes[10,11], Herman H H B M van Haagen[10], Elisabeth G E de Vries[1], Gerard J te Meerman[2], Cisca Wijmenga[2], Marcel A T M van Vugt[1] & Lude Franke[2]

Many cancer-associated somatic copy number alterations (SCNAs) are known. Currently, one of the challenges is to identify the molecular downstream effects of these variants. Although several SCNAs are known to change gene expression levels, it is not clear whether each individual SCNA affects gene expression. We reanalyzed 77,840 expression profiles and observed a limited set of 'transcriptional components' that describe well-known biology, explain the vast majority of variation in gene expression and enable us to predict the biological function of genes. On correcting expression profiles for these components, we observed that the residual expression levels (in 'functional genomic mRNA' profiling) correlated strongly with copy number. DNA copy number correlated positively with expression levels for 99% of all abundantly expressed human genes, indicating global gene dosage sensitivity. By applying this method to 16,172 patient-derived tumor samples, we replicated many loci with aberrant copy numbers and identified recurrently disrupted genes in genomically unstable cancers.

Thus far, thousands of genetic variants have been associated with cancer, other diseases and complex traits, and it has become clear that disease-associated SNPs and SCNAs (one of the hallmarks of cancer) often affect gene expression levels[1–4]. This observation indicates that changes in gene expression might be an important intermediate mechanism for genetic variants to exert their effect on the resulting phenotype. However, it is not clear to what extent the expression levels of genes are affected by SCNAs.

Here we aimed to systematically investigate the extent of gene dosage sensitivity. We hypothesized that every SCNA has an effect on gene expression levels but also that this effect is likely to be subtle and will generally be overshadowed by many other, non-genetic factors that have much stronger effects on gene expression (for example, physiological or metabolic factors).

To this end, we have developed a method called functional genomic mRNA profiling, or FGM profiling, that corrects gene expression data for major non-genetic factors. We first applied this method to gene expression data for which array comparative genomic hybridization (aCGH) data were available and observed a strong correlation between SCNAs and the FGM expression levels of genes that mapped within these SCNAs. We then identified a set of 16,172 unrelated patient-derived tumor samples and generated a comprehensive landscape of FGM profiles in these samples.

RESULTS

A regulatory model of the transcriptome

To identify the effects of genetic variation on gene expression levels, expression quantitative trait locus (eQTL) studies are typically employed. Often, principal-component analysis (PCA) is used to correct expression data for unwanted biological or technical effects[1,2]. PCA is usually performed on the expression data corresponding to the eQTL data set itself, resulting in a limited number of robust principal components. The strongest components (those explaining the largest proportion of variation in expression) are subsequently used to correct the expression data. Although these procedures have proven effective in identifying eQTLs, little attention has been paid so far to the makeup of these components. We hypothesized that many of these components reflect genuine non-genetic biological differences between samples, such as their metabolic or physiological states. We reasoned that a large compendium of gene expression data would permit the identification of many different biological phenomena. After identifying these phenomena, we should be able to quantify their presence for each individual sample in eQTL data sets and correct the expression data for non-genetic biological differences, resulting in increased power to identify eQTL effects.

[1]Department of Medical Oncology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [2]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [3]National Center for Advancing Translational Sciences, US National Institutes of Health, Rockville, Maryland, USA. [4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [5]Division of Endocrinology, Children's Hospital Boston, Boston, Massachusetts, USA. [6]Center for Basic and Translational Obesity Research, Children's Hospital Boston, Boston, Massachusetts, USA. [7]Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark. [8]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [9]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Haren, the Netherlands. [10]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [11]BioSemantics Group, Leiden Institute of Advanced Computer Science, Leiden University, Leiden, the Netherlands. [12]These authors contributed equally to this work. Correspondence should be addressed to R.S.N.F. (r.s.n.fehrmann@umcg.nl) or L.F. (lude@ludesign.nl)

Received 29 July 2014; accepted 2 December 2014; published online 12 January 2015; doi:10.1038/ng.3173

We therefore employed PCA on a large number of samples to identify a substantial set of robust principal components (which we henceforth refer to as transcriptional components, or TCs). We collected gene expression data for 77,840 samples from 4 Affymetrix platforms (33,427 human samples referred to as the human$_{large}$ data set, 17,309 human samples referred to as the human$_{small}$ data set, 17,081 mouse samples and 6,023 rat samples). Information on the cell types and tissues that these samples represent is provided in **Supplementary Tables 1** and **2**. The collection of samples was unbiased to ensure that we created a large and heterogeneous expression

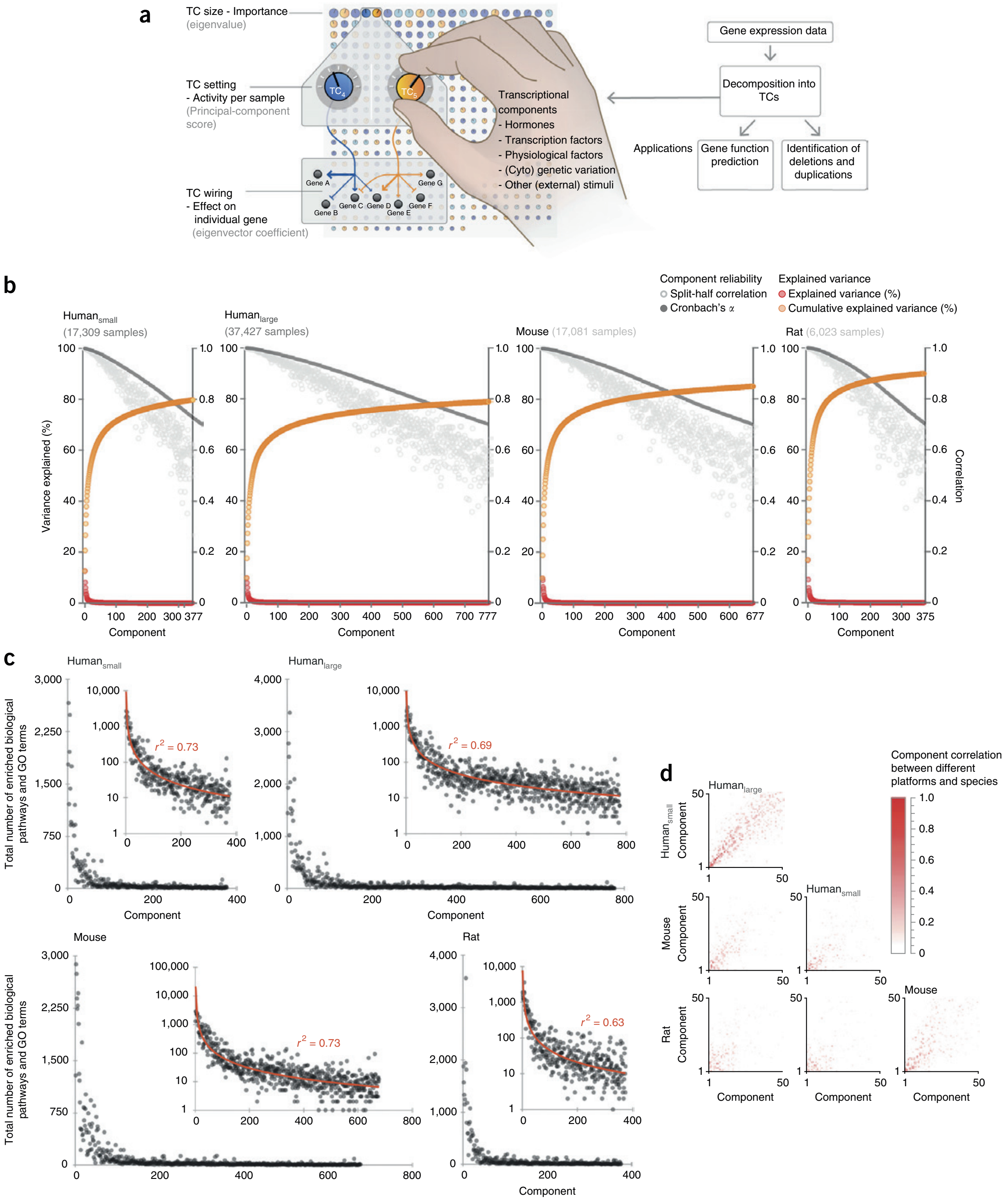


Figure 1 (continued)

data set in which many tissue types, disease states, cellular contexts, experimental variations, and therapeutic or chemical perturbations were present, thus maximizing our chances to identify different types of TCs.

Each TC explained a different part of the variance seen in gene expression in the 77,840 samples, and we hypothesized that these TCs reflect different expression-regulating factors (and corresponding cellular states) and thus can be seen as a regulatory model of the mRNA transcriptome (**Fig. 1a** and **Supplementary Tables 3–6**). The robustly estimated TCs (with Cronbach's $\alpha > 0.7$) captured 79–90% of the variation in gene expression, depending on the species or microarray platform (**Fig. 1b**). Gene set enrichment analysis (GSEA) with multiple pathway and gene set databases showed that all TCs had at least one functional gene set that was significantly enriched (at a permutation-based false discovery rate (FDR) of <0.05), indicating that the TCs capture genuine biological phenomena (**Fig. 1c** and **Supplementary Tables 7–10**). As expected, we observed that many of these components were conserved across the three species studied (**Fig. 1d**). PCA also enabled us to assess the activity of TCs in different cell types and tissues. An example of this is provided by the differential activity of TC_3 in brain tissues in comparison to cell lines (**Fig. 1e**). The higher activity (higher TC score; see the **Supplementary Note**) of TC_3 observed in brain samples clearly matched the GSEA results, in which genes in neuron-specific biological pathways were upregulated (**Fig. 1e**). In contrast, the lower activity of TC_3 observed in cell lines was associated with the upregulation of genes relevant to cell division and growth (**Fig. 1e**).

A TC-based gene network predicts biological functions

We hypothesized that genes with comparable biological functions are similarly regulated by TCs. Therefore, we constructed a TC-based gene network (publicly available; see URLs) with 19,997 genes

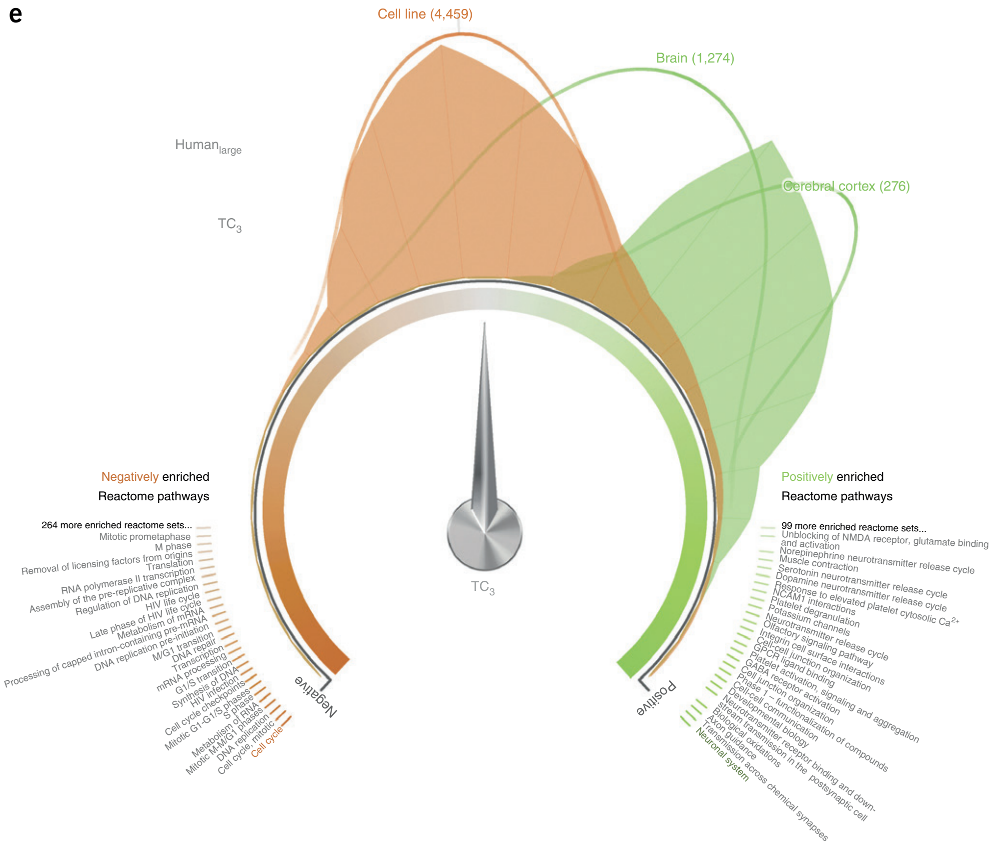


Figure 1 (continued). Transcriptional components (TCs). (**a**) PCA was used to define a regulatory model of the mRNA transcriptome, resulting in the identification of TCs. Each TC (eigenvector) captures a part of the variance (eigenvalue) seen in mRNA expression in our large, heterogeneous set of samples. Each TC is associated with an underlying factor regulating mRNA expression (and a corresponding cellular state). The TC (principal-component) score can be seen as an indirect measurement of the activity of the underlying regulatory factor in the sample under investigation. The sign of a probe set (eigenvector) coefficient in a TC defines the direction of the change in mRNA expression in relation to the TC score. We use the identified TCs to predict the functions of genes and to identify deletions and duplications in cancer samples. (**b**) Explained variance, Cronbach's α and split-half reliability for 377 human$_{small}$, 777 human$_{large}$, 677 mouse and 375 rat TCs. (**c**) GSEA on all TCs showed that all components were significantly enriched for at least one gene set. Insets show the same data on a $\log_{10}$ scale on the *y* axis. (**d**) Heat maps showing high concordance between the TCs identified in the different data sets. (**e**) Differential activity of TC_3 in brain tissue versus cell lines. Higher activity of TC_3 in brain showed upregulation of biological pathways relevant for neurons. In contrast, lower activity of TC_3 in cell lines was associated with upregulation of genes relevant for cell division and growth.

that enabled us to accurately predict gene function by using a 'guilt-by-association' procedure (see the **Supplementary Note** for details). Predictions were made on the basis of pathways and gene sets from Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) Reactome and Biocarta. These predictions outperformed an existing guilt-by-association approach using coexpression[5] (**Supplementary Fig. 1a** and **Supplementary Note**). In addition, we observed that genes with newly associated GO terms (including genes that were assigned to a GO term after the construction of our gene network) had significantly higher prediction *z* scores than genes that remained unassociated (**Supplementary Fig. 1b**), indicating that our predictions forecast new functions for previously unexplored genes.

In the context of genomic instability, we were interested in genes that had a similar cellular function as *BRCA1* and *BRCA2*, two key genes involved in homologous recombination–mediated DNA repair[6]. Such genes could potentially be used to identify homologous recombination–defective tumor cells or could be exploited as therapeutic targets to enhance the sensitivity of cancer cells that are otherwise proficient in homologous recombination to DNA-damaging agents. On the basis of the gene network, we ranked all genes according to their degree of coregulation with *BRCA1* and *BRCA2* (**Fig. 2a**). We could readily identify genes known to be involved in homologous recombination (for example, *RAD51*, *RAD51AP1*, *EXO1*, *BLM* and *MRE11A*; **Fig. 2a**). In addition, many genes predicted to be coregulated with *BRCA1* and *BRCA2* function in the initiation and progression of replication forks, in good agreement with *BRCA1* and *BRCA2* being required for the repair of replication-induced DNA lesions. *FEN1* (encoding flap endonuclease-1) showed the highest degree of coregulation with *BRCA1* and *BRCA2* (**Fig. 2a**). FEN1 regulates DNA replication through its role in Okazaki fragment processing and

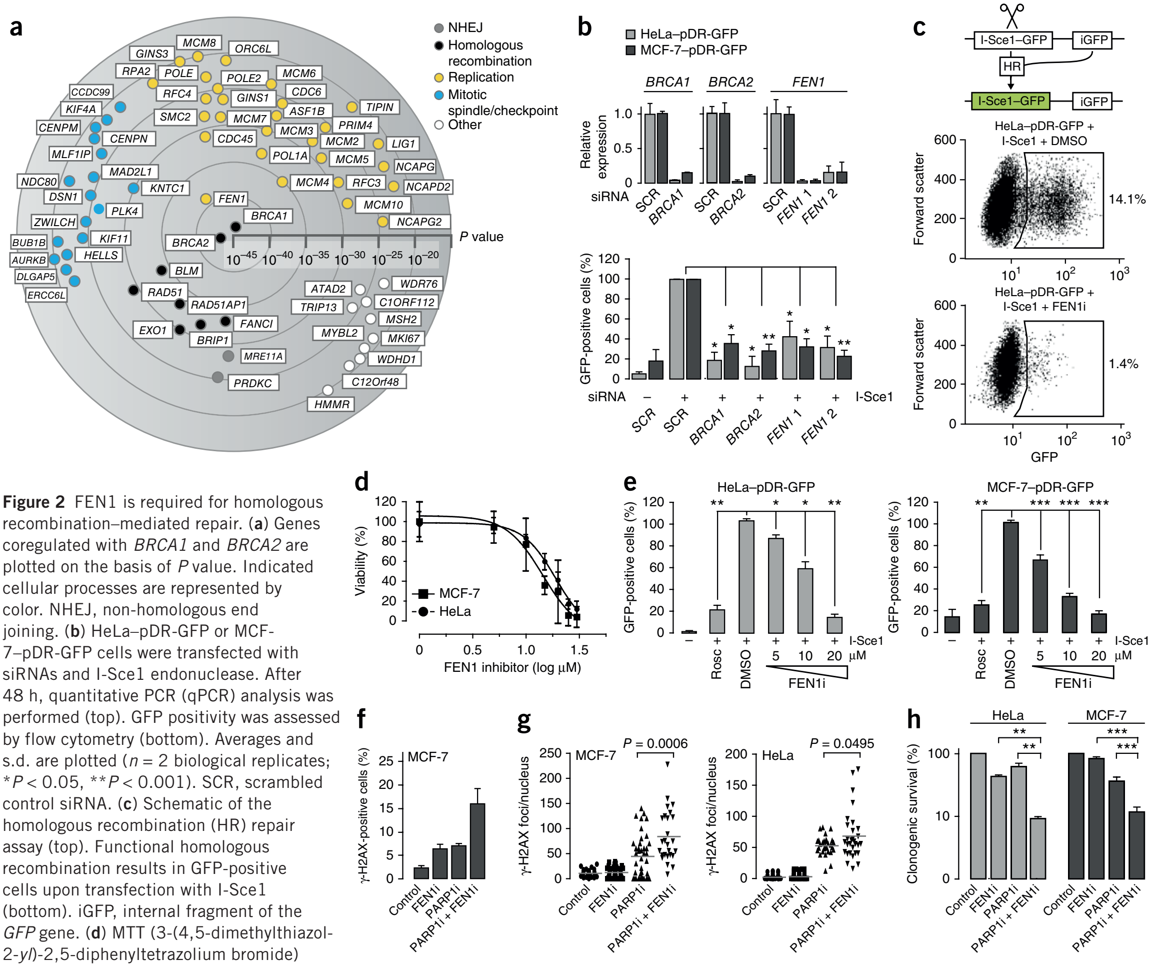


Figure 2 FEN1 is required for homologous recombination–mediated repair. (**a**) Genes coregulated with *BRCA1* and *BRCA2* are plotted on the basis of *P* value. Indicated cellular processes are represented by color. NHEJ, non-homologous end joining. (**b**) HeLa–pDR-GFP or MCF-7–pDR-GFP cells were transfected with siRNAs and I-Sce1 endonuclease. After 48 h, quantitative PCR (qPCR) analysis was performed (top). GFP positivity was assessed by flow cytometry (bottom). Averages and s.d. are plotted ($n = 2$ biological replicates; $*P < 0.05$, $**P < 0.001$). SCR, scrambled control siRNA. (**c**) Schematic of the homologous recombination (HR) repair assay (top). Functional homologous recombination results in GFP-positive cells upon transfection with I-Sce1 (bottom). iGFP, internal fragment of the *GFP* gene. (**d**) MTT (3-(4,5-dimethylthiazol-2-*yl*)-2,5-diphenyltetrazolium bromide) viability assays using MCF-7 and HeLa cells after FEN1 inhibitor treatment. Averages and s.d. are shown ($n = 2$ biological replicates). (**e**) HeLa–pDR-GFP or MCF-7–pDR-GFP cells were transfected with I-Sce1 and treated with FEN1 inhibitor (FEN1i) or roscovitin (Rosc; 25 μM) as a positive control. After 48 h, GFP positivity was assessed by flow cytometry. Averages and s.d. are indicated ($n = 2$ biological replicates; $*P < 0.05$, $**P < 0.001$, $***P < 0.001$). (**f**) MCF-7 cells were treated with the PARP1 inhibitor ABT-888 (PARP1i; 10 μM) and with FEN1 inhibitor (10 μM). After 24 h, cells were stained for γ-H2AX (a marker of DNA double-strand breaks) and analyzed by flow cytometry. Averages and s.d. are indicated ($n = 2$ biological replicates). (**g**) MCF-7 and HeLa cells were treated as in **f** and analyzed by fluorescence microscopy (γ-H2AX foci of >25 nuclei per condition were counted). Gray bars represent median values. (**h**) MCF-7 and HeLa cells were grown in the presence of ABT-888 (10 μM) and FEN1 inhibitor (5 μM). After 14 d, surviving colonies were counted. Averages and s.d. are plotted ($n = 2$ biological replicates; $**P < 0.001$, $***P < 0.001$).

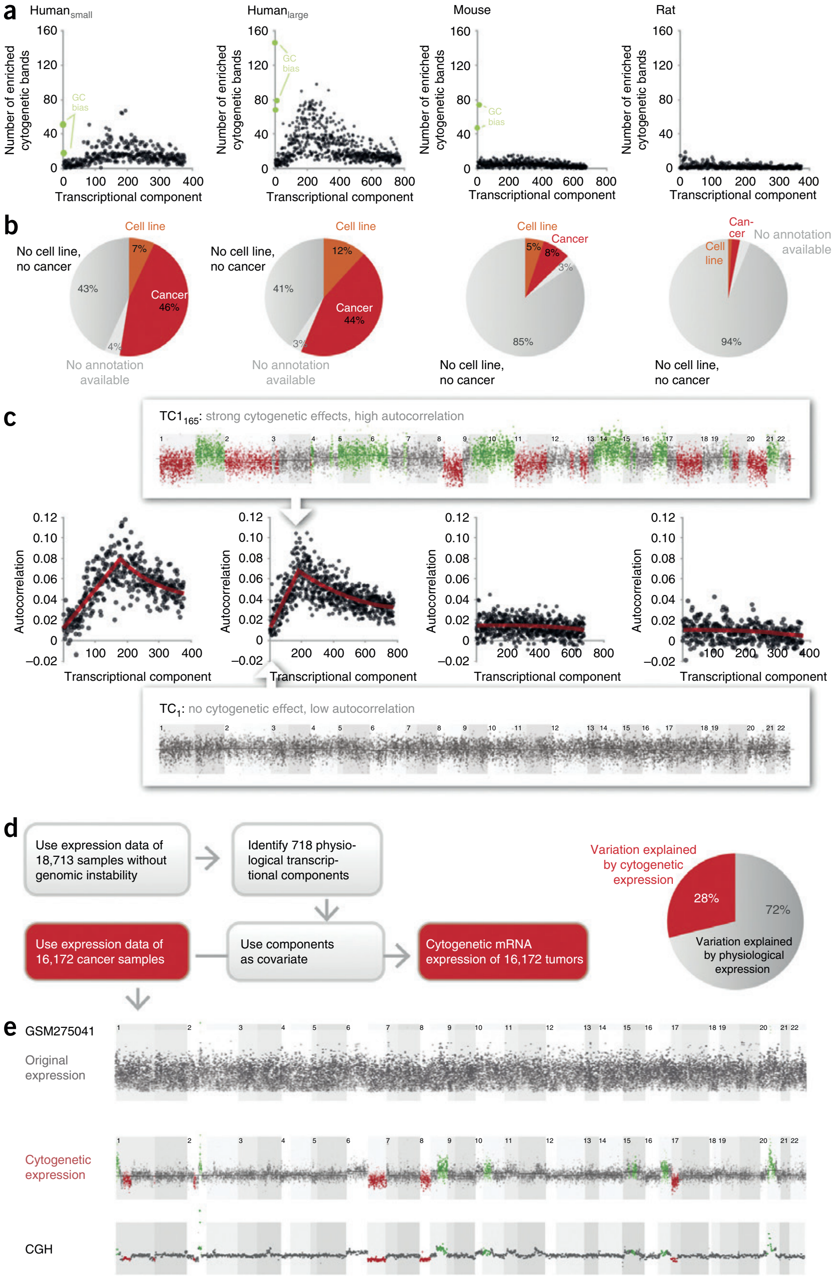


Figure 3 FGM expression profiles. (**a**) A subset of human TCs were highly enriched for gene sets containing genes that map to the same chromosome band. (**b**) This observation occurred solely in the human data sets, as approximately half of the samples were derived from cancer tissue or cancer cell lines, harboring many SCNAs. (**c**) To explore the hypothesis that specific TCs represent SCNAs, we calculated the autocorrelation per TC. We observed a near-linear increase in autocorrelation with TC number until approximately TC$_{165}$ in the human$_{large}$ set. The same pattern was observed in the human$_{small}$ data set. Individual visualization of the TCs that showed high autocorrelation demonstrated that, for most of these specific TCs, nearly all genes in specific genomic regions showed either increased or decreased expression. Insets show probe set coefficients plotted according to their genomic positions for a TC with strong cytogenetic effects and a TC with no cytogenetic effects. (**d**) On the basis of these results, we developed a computational framework in which we could infer cytogenetic profiles from a standard mRNA expression profile using TCs. (**e**) FGM expression profiles are a good proxy of copy number variation. Small deletions and duplications, affecting only specific cytobands, could be readily identified.

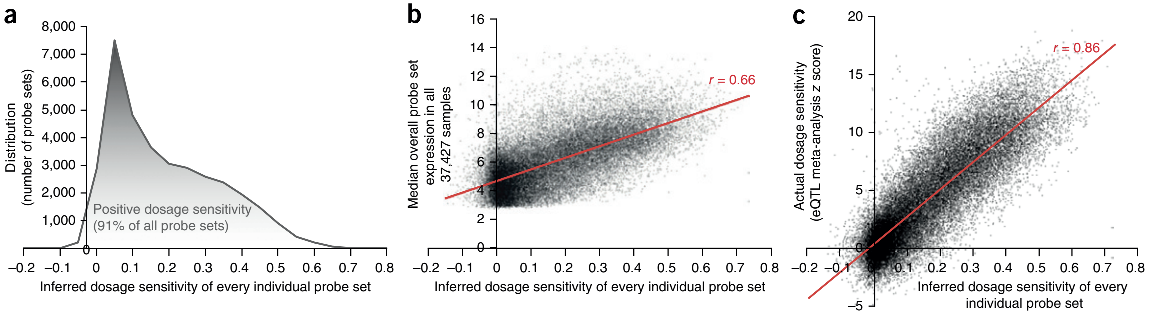


Figure 4 The vast majority of genes are dosage sensitive. (**a**) Distribution of correlations over all samples between the FGM expression of individual probe sets and the average FGM expression of the chromosome arm to which these probe sets map. We observed that 91% of all autosomal probe sets correlated positively with average FGM expression, indicating that nearly all genes are dosage sensitive. (**b**) A bimodal distribution in the extent of dosage sensitivity was observed in the distribution of correlations that could be explained by the overall median expression levels of probe sets in all samples (Pearson $r = 0.66$). (**c**) On the level of individual probe sets, we observed a strong relationship (Pearson $r = 0.85$) between the predicted and empirically observed dosage sensitivities in the eQTL meta-analysis of 470 samples.

base-excision repair (BER)[7], and it is highly expressed in proliferating tissues. In addition, mutation of *FEN1* has been causally linked to tumor initiation[8]. FEN1 has only been indirectly linked to homologous recombination in lower eukaryotes, and no role for FEN1 in homologous recombination in mammalian cells has been identified so far[9]. We assessed the function of FEN1 in the process of homologous recombination by inactivating *FEN1* by small interfering RNA (siRNA) (**Fig. 2b**), as well as by chemical inhibition of FEN1 (**Fig. 2c**,**d**). Inactivation of FEN1 clearly impaired homologous recombination–mediated repair, as assessed with a stably integrated GFP-based homologous recombination substrate[10], in both human breast and cervical cancer cells (**Fig. 2b**,**c**,**e**). Moreover, we observed increased numbers of DNA breaks when FEN1 inhibition was combined with PARP1 inhibition (**Fig. 2f**,**g**). Notably, FEN1 inhibition resulted in increased sensitivity to PARP1 inhibition (**Fig. 2h**), as reported for other homologous recombination–deficient tumor cells[11,12].

In addition, our gene network predicted *SMIM1* to function in the 'hemoglobin metabolic process'. Through subsequent experimental validation of this prediction, we showed that *SMIM1* underlies the Vel blood group[13]. We also showed that genes that map in genetic loci associated with educational attainment have a clear neuronal function[14].

Some TCs describe the effects of genetic variants

Once we had established that the identified TCs described biologically coherent phenomena, we aimed to use them to correct gene expression levels to thereby increase our statistical power to identify the eQTL effects of SCNAs (shown to be effective earlier[2]). However, we observed that a subset of the human TCs (each capturing only a small proportion of the total variation in expression) strongly affected the expression of all genes in a specific cytogenetic band (**Fig. 3a**), which we hypothesized to result from large SCNAs that were present in approximately half of the human samples that had been annotated as cancer tissues or cancer cell lines (in contrast to the mouse and rat data, which contained very few cancer samples) (**Fig. 3b**). To discriminate between genomically stable and unstable samples, we investigated this genomic colocalization of expression effects in more detail (using autocorrelation; **Fig. 3c** and **Supplementary Note**). We observed that the first TCs (TC$_1$ to TC$_{50}$, which captured most expression variation) hardly showed this genomic colocalization of expression effects. The other TCs showed a near-linear increase in autocorrelation with increasing TC number, up to TC$_{165}$ in the human$_{\text{large}}$ set, where autocorrelation was maximal (indicating strong colocalization of expression effects; **Fig. 3c**).

On the basis of these observations, we could define a subset of 18,713 samples in the human$_{\text{large}}$ data set that showed no evidence of genomic instability (no evident SCNAs; **Supplementary Note**). We subsequently performed PCA again on this subset of 18,713 samples and identified 718 robustly estimated non-genetic TCs. We used these 718 non-genetic TCs as covariates to correct the expression data for the other 18,714 human$_{\text{large}}$ samples that did show evidence of genomic instability (**Fig. 3d**; details provided in the **Supplementary Note**) and observed that the residual expression signal (explaining, on average, only 28% of the total expression variation) was strongly correlated with copy number variation (**Fig. 3e**).

Residual expression levels reflect SCNAs

We hypothesized that this residual expression signal (hereafter referred to as the functional genomic mRNA profile, or FGM profile) accurately reflects the presence of deletions and duplications. We first investigated 20 mRNA expression profiles for samples with known trisomy. When employing our method, each trisomy was clearly visible: nearly all genes on these trisomic chromosomes showed increased FGM expression (**Supplementary Fig. 2a**). The trisomies were clearly visible in both human leukemic blasts and fibroblasts, indicating that the FGM expression profiles were stable across different cell types (**Supplementary Fig. 2b**,**c**).

To validate our method on samples with smaller SCNAs, we performed an eQTL meta-analysis on expression data from 470 samples (51 *HER2* (*ERBB2*)-amplified breast cancers, 173 inflammatory breast cancers and 246 multiple myelomas) for which aCGH data were also available (**Supplementary Note**). We observed a high concordance between aCGH data and FGM expression (median Pearson $r = 0.64$; **Supplementary Fig. 2d** and **Supplementary Note**). The FGM expression levels for 85.7% of all the autosomal probe sets correlated positively with the aCGH data. Estimated sensitivity and specificity (based on the 470 samples described above) for the detection of copy number gains and losses for different genomic lengths are provided in **Supplementary Table 11**. See the **Supplementary Note** and **Supplementary Figures 3–7** for comparisons with previous expression-based methods for SCNA detection. Software to infer FGM expression from standard expression profiles is publicly available (see URLs).

The vast majority of genes are dosage sensitive

Although 85.7% of the tested probe sets positively correlated with aCGH gene dosage on the basis of the eQTL meta-analysis in 470 samples, this proportion is likely to be an underestimate of the total percentage of genes that are dosage sensitive, as it is only possible to identify a significant correlation between copy number and gene expression when variation in copy number is actually present for a given genomic locus.

To investigate dosage sensitivity in more detail, we assumed that our large data set represented a library of many SCNAs of different sizes. A subset of these SCNAs reflected entire chromosome arms. We therefore hypothesized that, if we were to analyze the FGM expression data by chromosome arm, the most dominant pattern would reflect a deletion or amplification of the entire chromosomal arm. To investigate this hypothesis, we performed PCA on the FGM expression data for each chromosome arm over all samples. For each of the chromosome arms, the first principal component (PC_1), representing the most dominant FGM expression pattern, indeed described a complete duplication or deletion of that arm (**Supplementary Fig. 8**). The correlation of an individual probe set (factor loading) with PC_1 indicates

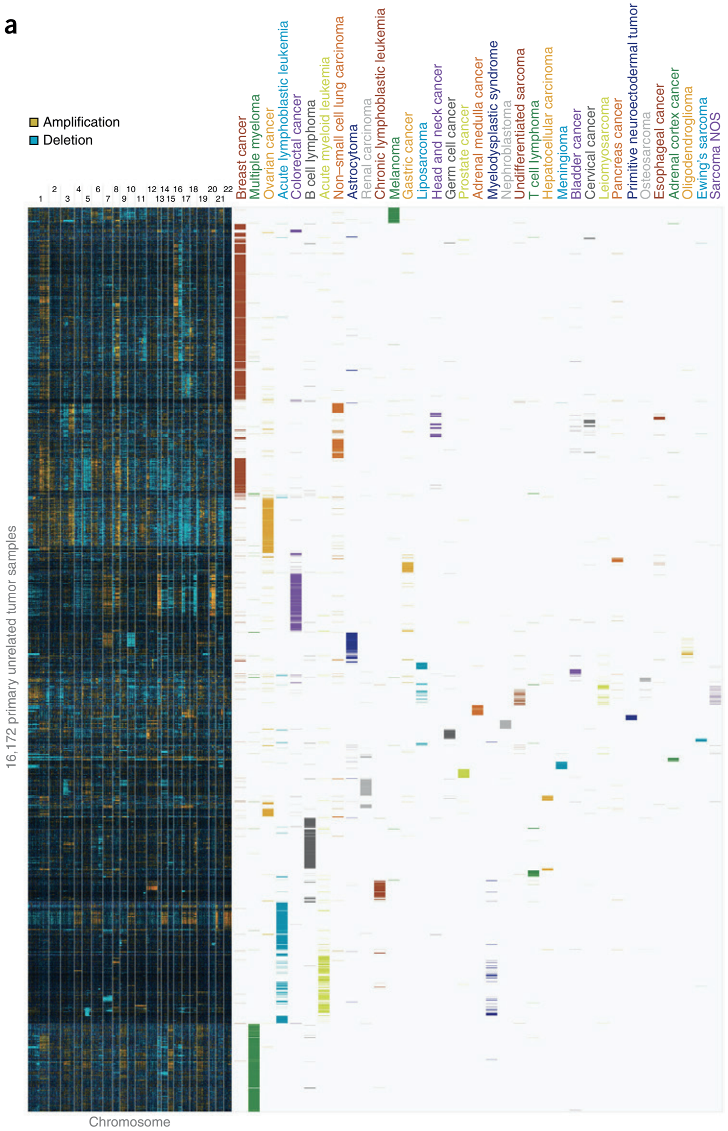
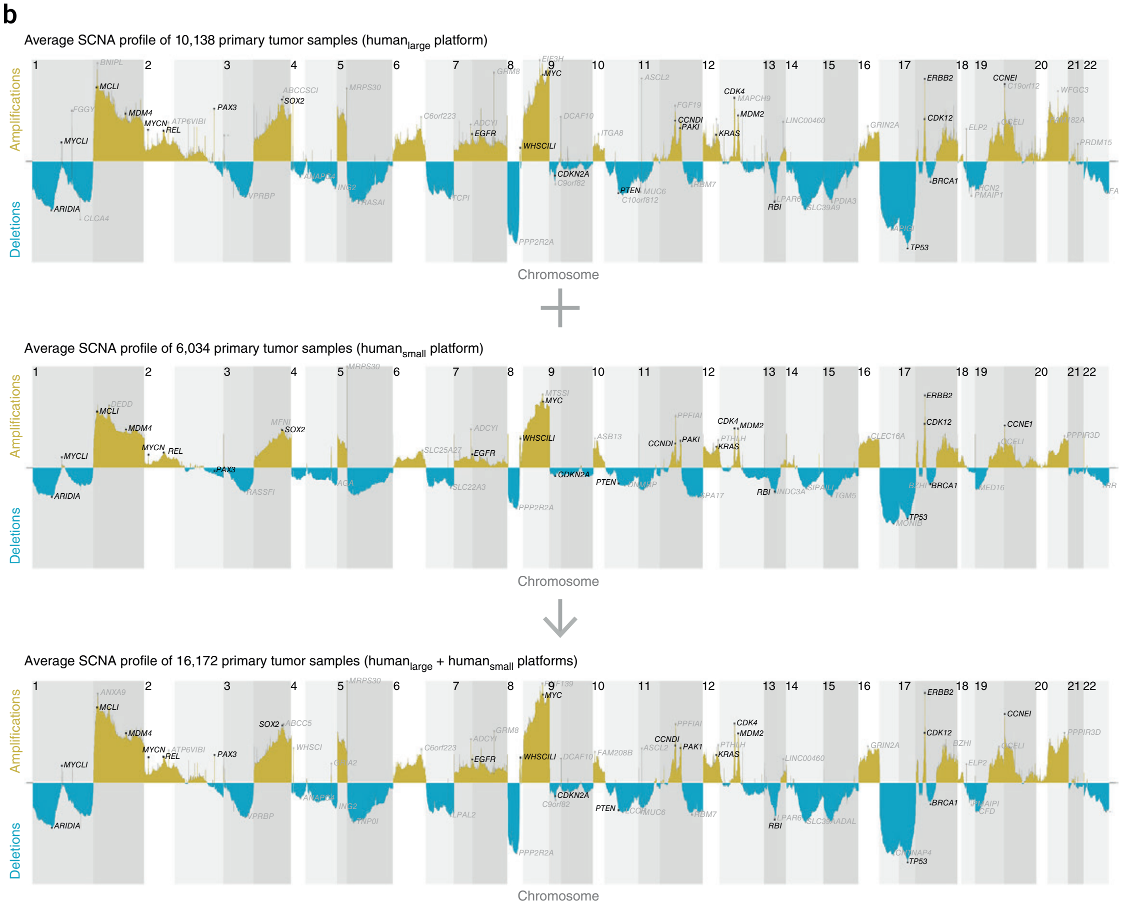


Figure 5 The FGM landscape in 16,172 unrelated patient-derived cancer samples. (**a**) We applied our method to 16,172 unrelated tumor samples from patients to determine the FGM landscape. We applied hierarchical clustering (average linkage, uncentered correlation) to define molecular subtypes. This analysis identified clear commonalities and differences between different types of cancer. (**b**) The frequently used circular binary segmentation algorithm was applied to the FGM profiles of each of the 16,172 tumor samples (10,138 samples on the U133 Plus 2.0 array platform and 6,034 samples on the U133A platform) to define regions with abnormal copy numbers. Subsequently, the average SCNA profile was defined across these 16,172 samples.



Average SCNA profile of 10,138 primary tumor samples (human$_{large}$ platform)

Average SCNA profile of 6,034 primary tumor samples (human$_{small}$ platform)

Average SCNA profile of 16,172 primary tumor samples (human$_{large}$ + human$_{small}$ platforms)

Figure 5 (continued)

the extent of dosage sensitivity. Of all the autosomal probe sets, 91% were dosage sensitive to some degree: that is, they had a positive correlation with PC$_1$ (**Fig. 4a**).

Genes that were generally abundantly expressed were typically more dosage sensitive than genes that showed low overall expression levels (Pearson $r = 0.7$, $P < 1.0 \times 10^{-300}$; **Fig. 4b**): when taking the top 25% most highly expressed probe sets across all samples, the mean dosage sensitivity (factor loading) was 0.32, and 98.9% of these probe sets showed positive dosage sensitivity. This indicates that, if a gene is abundantly expressed (and its expression thus can be accurately quantified using microarrays), it nearly always is dosage sensitive.

This predicted dosage sensitivity for the probe sets was very similar to the empirically observed dosage sensitivity in the eQTL meta-analysis (Pearson's $R = 0.86$ between the factor loadings and the aCGH eQTL meta-analysis z scores; **Fig. 4c**).

The FGM landscape in 16,172 cancer samples

FGM profiling is particularly useful because of the public availability of microarray expression profiles for thousands of cancer samples.

We exploited data on the genetic variation occurring in these tumor samples in combination with our method to define an FGM landscape. Analysis was confined to unrelated tumor samples from patients in the combined human$_{large}$ and human$_{small}$ data sets. To identify these tumor samples, we first developed a method (see the **Supplementary Note** for details) to exclude cell line samples. This method accurately distinguished between cell line samples and all other samples (Wilcoxon Mann-Whitney U test, $P < 1.0 \times 10^{-300}$, area under the curve (AUC) = 0.9942). To exclude duplicate samples, we then developed a method (see the **Supplementary Note** for details) to accurately detect and exclude genetically identical samples in expression data, even if two different cell types or tissues had been assayed for a single individual. After manual curation of all remaining samples, 16,172 unrelated tumor samples from patients were included for further analyses (see **Supplementary Table 12** for the distribution among the 41 tumor subtypes present in this data set). We then applied our method to these samples to determine the FGM landscape (**Fig. 5a**).

We applied hierarchical clustering to define molecular subtypes, which identified clear commonalities and differences between different types of cancer. Highly altered FGM expression, which was confined to specific chromosomal regions, was observed between and across different tumor types (**Fig. 5a**). This illustrates that different cancer subtypes share genetic alterations driving biological pathways relevant for their tumor behavior and treatment response.

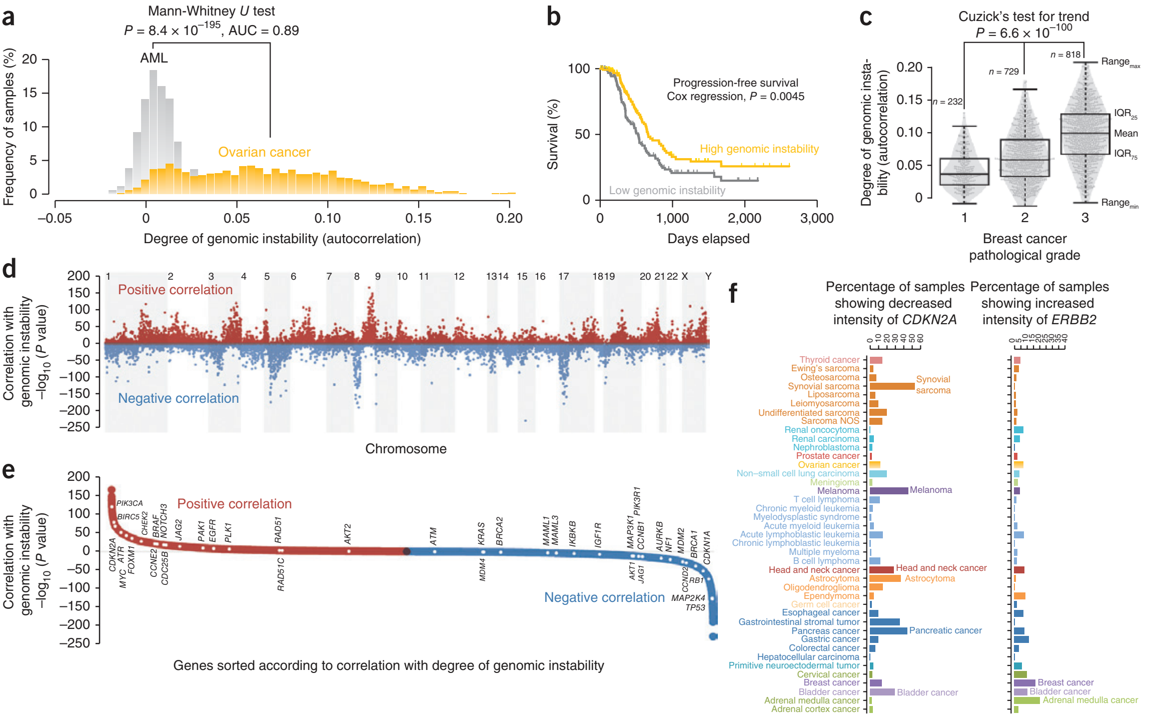


Figure 6 Associations with genomic instability. (**a**) A clear shift toward a higher degree of genomic instability was observed for ovarian cancer (n = 1,255) in comparison to AML (n = 1,540). (**b**) A significant association was found between a higher degree of genomic instability and better progression-free survival in 412 primary, late-stage, high-grade serous tumor samples from TCGA. (**c**) Within a set of 1,779 breast cancer samples, we observed a clear positive association between the degree of genomic instability and tumor grade. IQR, interquartile range. (**d**) Manhattan plot for genome-wide associations of FGM expression with genomic instability. (**e**) Sorted associations of FGM expression with genomic instability. (**f**) Percentage of significantly increased and decreased signals across 41 tumor types for *CDKN2A* and *ERBB2*, respectively.

We then applied the frequently used DNACopy software to define regions with abnormal copy number in the FGM profiles of each of the 16,172 tumor samples. Specifically, we analyzed what the average SCNA profile was across these 16,172 samples (**Fig. 5b**). We observed several narrow peaks that typically only comprised a few genes. Several of these genes are well known as being frequently altered in cancer (such as *TP53*, *ERBB2*, *CCNE1*, *MYC*, *CDK4*, *CDK12*, *MDM2*, *PAX3* and *RB1*). These results indicate that our method is able to identify amplifications and deletions that can be as small as an individual gene (for example, *PAX3*) and it provides insight into the genes that are most often genetically altered in cancer. To further highlight the relevance of this FGM landscape, we performed a genome-wide association analysis between the expression of individual genes (FGM expression) and the degree of genomic instability in 15,204 of the 16,172 tumor samples (samples were excluded if they were part of a meta-analysis batch with fewer than 10 samples; see the **Supplementary Note** for details). We reasoned that normal cells do not tolerate high levels of genomic instability[15], and it is therefore remarkable that certain tumor cells can cope with these levels of instability. Insight into how these tumor cells are rewired to survive genomic instability might uncover new therapeutic targets and will be important to help improve the treatment outcome for individuals with genomically unstable cancers.

For each tumor sample, we determined the degree of genomic instability. We observed a clear shift (Mann-Whitney U test, $P = 8.4 \times 10^{-195}$, AUC = 0.89) toward a higher degree of genomic instability for ovarian cancer (n = 1,255) in comparison to acute myeloid leukemia (AML; n = 1,540; **Fig. 6a**). Our findings coincide with the recent classification of mutation-driven (M class; including AML) versus copy number–driven (C class; including ovarian cancer) cancers in The Cancer Genome Atlas (TCGA) data sets[4].

Recent evidence suggests that a higher degree of genomic instability leads to increased sensitivity to certain DNA-damaging chemotherapeutics. Currently, debulking surgery followed by DNA-damaging platinum-based chemotherapy is considered to be standard care for serous ovarian cancer. We performed a meta-analysis on 842 primary late-stage, high-grade serous tumor samples (**Supplementary Note**) and found that a higher degree of genomic instability was associated with a better overall survival rate (Cox regression, $P = 0.0014$). In addition, when we analyzed 412 primary late-stage, high-grade serous tumor samples from TCGA[16], we found significant association between a higher degree of genomic instability and longer progression-free survival (Cox regression, $P = 0.0045$; **Fig. 6b**).

A second prototypical tumor type with a high degree of genomic instability is triple-negative breast cancer (TNBC)[17]. As expected, there was a clear shift (Mann-Whitney U test, $P = 1.6 \times 10^{-101}$, AUC = 0.24) toward a higher degree of genomic instability for TNBCs (n = 672) in comparison to non-TNBCs (n = 3,053) (**Supplementary Fig. 9**). For all breast cancer samples, we also observed a clear positive association between the degree of genomic instability and pathological grade (Cuzick's trend, $P = 6.6 \times 10^{-100}$; **Fig. 6c**), which agrees well with high-grade tumors often having elevated levels of genomic instability.

Meta-analysis association results for genome-wide association of gene expression with genomic instability are provided in **Figure 6d**,e and **Supplementary Table 13**. When we compared our findings with the TCGA analysis of prototypical genomically unstable cancers (TNBC and serous ovarian cancer), it was reassuring to find many overlapping genes. Genes that are negatively associated with genomic instability (including *TP53*, *CDKN2A*, *RB1*, *BRCA1*, *BRCA2* and *ATM*) were frequently found to be inactivated in TNBCs and serous ovarian cancers. Conversely, genes that are positively associated with genomic instability were frequently found to be amplified (including *MYC*, *CCNE1*, *PIK3CA* and *BIRC5*)[16,18]. Further research is required to uncover the mechanistic relationship of these genes with genome maintenance.

In addition, we determined for each individual gene across 41 tumor types the percentage of samples in the FGM landscape with a significantly increased or decreased signal (**Supplementary Tables 14** and **15**). Such altered signals could represent underlying SCNAs (**Fig. 3e** and **Supplementary Fig. 2**). The resulting percentages can be used to assess the potential relevance of new and existing therapeutic targets for the various cancer subtypes. Two examples show how this approach can be used to discover cancer-relevant genomic alterations (**Fig. 6f**). The signal for the tumor-suppressor gene *CDKN2A*, for instance, was shown to be significantly decreased in samples of bladder cancer (29%), pancreatic cancer (44%), head and neck cancer (28%) and melanoma (45%), findings that are in good agreement with previous reports[19–22] and with recent large-scale genomic interrogation of cancer subtypes[23,24]. Conversely, we could readily identify amplified signals for the *ERBB2* gene in breast cancer (17%), bladder cancer (10%) and adrenal medulla cancer (20%), again in line with previous observations[25–27].

DISCUSSION

In this study, we reanalyzed the expression profiles of 77,840 samples and observed that a limited number of TCs capture the majority of variation that is present in the mRNA transcriptome. On the basis of these TCs, we have developed a method to infer the likely biological function of candidate genes by implementing a guilt-by-association approach; the tool predicts likely functions on the basis of gene coregulation. Our approach is conceptually different from methods that rely on coexpression: typical guilt-by-association approaches to predict the function of genes use the over-representation of known functions among genes that are coexpressed with the gene of interest to predict its likely function[28]. However, when one is clustering genes on the basis of coexpression, usually only a few clusters will become apparent, which each comprise many strongly coexpressed genes (attributable to a limited number of biological phenomena that each have very strong effects on gene expression levels). Here we chose to use PCA rather than coexpression analysis. While we did capture the strong biological phenomena with a few individual TCs, we also captured many phenomena that were much more subtle. Our gene function prediction algorithm treats each of the TCs equally and can therefore extract more meaningful information from gene expression data to infer the likely function of genes. In other words, our approach relates genes with similar transcriptional regulation (similar wiring) rather than those with similar coexpression patterns (the end product of transcriptional control).

When we corrected the gene expression data of individuals for non-genetic TCs, we observed that the residual gene expression levels correlated strongly with copy number variation. We estimate that the expression of 98.9% of all abundantly expressed human genes correlates positively with copy number to some degree, The realization that copy number variation has a consistent effect on gene expression levels has several implications, as detailed below.

Previously, extensive gene dosage sensitivity was suggested in whole-chromosome duplication syndromes (for example, Down, Turner and Patau syndromes) or experimentally established trisomies, in which gene expression over the entire duplicated chromosome appeared to be upregulated[29]. Our data suggest that this is not a chromosome-specific or cell type–specific effect, explaining why aneuploidy invariably leads to unbalanced gene expression levels and consequent proteotoxic stress[30].

In addition, several integrative cancer studies have concentrated only on those genes whose expression levels significantly correlate with SCNAs[4]. This approach might bias researchers toward considering only genes whose expression levels are not strongly regulated by regulatory processes. It is conceivable that reanalysis of these expression data using the method we propose here (correcting the data for all non-genetic regulatory processes) will show that the expression levels of nearly all genes in these SCNAs correlate with copy number, making every gene within these SCNAs equally interesting for follow-up investigation. In addition, we speculate that genes that are typically under the strong control of many non-genetic regulatory factors are likely to have crucial functions and that they may represent more interesting candidate driver genes for disease than genes whose expression levels are predominantly affected by SCNAs. Furthermore, gene expression microarray technology is still a widely used technique to identify molecular cancer subtypes. However, there is limited overlap between published gene expression profiles associated with distinct tumor behavior or treatment response[31]. This low reproducibility might, in part, be due to the influence of many non-genetic factors on gene expression, including factors that may not be relevant for the observed tumor behavior.

Because gene expression data have now been generated for hundreds of thousands of samples (available in public repositories such as the Gene Expression Omnibus, GEO), our FGM profiling method will enable the reinterpretation of many (cancer) studies, as our method can be used to infer SCNAs and their downstream effect on gene expression levels and allows for an immediate correlation with various clinical outcome measures. However, as there are few clinicopathological data publicly available thus far, we challenge researchers to apply our method to their microarray samples and associate the resulting FGM profiles with clinicopathological variables to evaluate how such profiles can predict tumor behavior or therapy response, for instance. We expect that such reanalyses might well provide new biological insights in the near future.

In this study, we show the advantages and potential relevance of FGM profiling in combination with large-scale ('big data') analysis. Our meta-analysis on patient-derived tumor samples provided new insights into the rewiring of genomically unstable tumors that could make them dependent on specific genes or biological pathways for their survival, thereby identifying potentially synthetic lethal targets in a genomically unstable context. In addition, the predicted genomic amplifications and deletions we have described, on the basis of the FGM landscape in 16,172 tumors, can already be used to assess the potential relevance of new and existing therapeutic targets for various cancer subtypes.

URLs. The TC-based gene network, FGM profiler, and additional data sets and results are accessible at http://www.genenetwork.nl/fgmp/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J.L. Senior for editing the manuscript. This work was financially supported by grants from the Netherlands Organization for Scientific Research (NWO-VENI grant 916-10135 to L.F., NWO VIDI grant 916-76062 to M.A.T.M.v.V. and NWO VIDI grant 917-14374 to L.F.), a Horizon Breakthrough grant from the Netherlands Genomics Initiative (grant 92519031 to L.F.), a grant from the Van der Meer–Boerema Foundation to M.K. and grants from the Dutch Cancer Society (RUG 2011-5093 to M.A.T.M.v.V. and RUG 2013-5960 to R.S.N.F.). In addition, this study was financed in part by the SIA-raakPRO subsidy for project BioCOMP. The research leading to these results has received funding from the European Community's Health Seventh Framework Programme (FP7/2007–2013) under grant agreement 259867. This publication was made possible through the support of a grant from the John Templeton Foundation. Additional support for E.A.S. was provided by the Leiden Centre for Data Sciences.

AUTHOR CONTRIBUTIONS

R.S.N.F. and L.F. conceived this study. M.K. performed the *in vitro* experiments. D.M. and A.S. synthesized chemical compounds. R.S.N.F., J.M.K., T.H.P., J.N.H., R.C.J., E.A.S., H.H.H.B.M.v.H., H.-J.W., G.J.t.M., M.A.T.M.v.V. and L.F. performed analyses. R.S.N.F., J.M.K., T.H.P., E.G.E.d.V., C.W., M.A.T.M.v.V. and L.F. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Fehrmann, R.S.N. *et al.* *Trans*-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011).
2. Westra, H.-J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
3. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
4. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
5. Wolfe, C.J., Kohane, I.S. & Butte, A.J. Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks. *BMC Bioinformatics* **6**, 227 (2005).
6. Moynahan, M.E. & Jasin, M. Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nat. Rev. Mol. Cell Biol.* **11**, 196–207 (2010).
7. Tomlinson, C.G., Atack, J.M., Chapados, B., Tainer, J.A. & Grasby, J.A. Substrate recognition and catalysis by flap endonucleases and related enzymes. *Biochem. Soc. Trans.* **38**, 433–437 (2010).
8. Zheng, L. *et al.* *Fen1* mutations result in autoimmunity, chronic inflammation and cancers. *Nat. Med.* **13**, 812–819 (2007).
9. Kikuchi, K. *et al.* Fen-1 facilitates homologous recombination by removing divergent sequences at DNA break ends. *Mol. Cell. Biol.* **25**, 6948–6955 (2005).
10. Pierce, A.J., Johnson, R.D., Thompson, L.H. & Jasin, M. XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. *Genes Dev.* **13**, 2633–2638 (1999).
11. Farmer, H. *et al.* Targeting the DNA repair defect in *BRCA* mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).
12. Bryant, H.E. *et al.* Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, 913–917 (2005).
13. Cvejic, A. *et al.* *SMIM1* underlies the Vel blood group and influences red blood cell traits. *Nat. Genet.* **45**, 542–545 (2013).
14. Rietveld, C.A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).
15. Hakem, R., de la Pompa, J.L. & Mak, T.W. Developmental studies of *Brca1* and *Brca2* knock-out mice. *J. Mammary Gland Biol. Neoplasia* **3**, 431–445 (1998).
16. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
17. Shah, S.P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
18. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
19. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
20. Caldas, C. *et al.* Frequent somatic mutations and homozygous deletions of the p16 (*MTS1*) gene in pancreatic adenocarcinoma. *Nat. Genet.* **8**, 27–32 (1994).
21. Monzon, J. *et al.* *CDKN2A* mutations in multiple primary melanomas. *N. Engl. J. Med.* **338**, 879–887 (1998).
22. Williamson, M.P., Elder, P.A., Shaw, M.E., Devlin, J. & Knowles, M.A. *p16* (*CDKN2*) is a major deletion target at 9p21 in bladder cancer. *Hum. Mol. Genet.* **4**, 1569–1577 (1995).
23. Santarius, T., Shipley, J., Brewer, D., Stratton, M.R. & Cooper, C.S. A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer* **10**, 59–64 (2010).
24. Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
25. Hovey, R.M. *et al.* Genetic alterations in primary bladder cancers and their metastases. *Cancer Res.* **58**, 3555–3560 (1998).
26. Slamon, D.J. *et al.* Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* **244**, 707–712 (1989).
27. Sworczak, K. *et al.* Gene copy numbers of erbB oncogenes in human pheochromocytoma. *Oncol. Rep.* **9**, 1373–1378 (2002).
28. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
29. Torres, E.M. *et al.* Identification of aneuploidy-tolerating mutations. *Cell* **143**, 71–83 (2010).
30. Oromendia, A.B., Dodgson, S.E. & Amon, A. Aneuploidy causes proteotoxic stress in yeast. *Genes Dev.* **26**, 2696–2708 (2012).
31. Fan, C. *et al.* Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.* **355**, 560–569 (2006).

ONLINE METHODS

Data acquisition. Publicly available raw microarray expression data for three different species (*Homo sapiens*, *Mus musculus* and *Rattus norvegicus*) were collected from GEO[32]. Preprocessing and aggregation of raw data were performed according to the robust multi-array average (RMA) algorithm in combination with quantile normalization. PCA quality control of the resulting expression data was performed as previously described[33,34]. Samples were annotated with MeSH terms using a text-mining algorithm (described in more detail in the **Supplementary Note**). We used a conservative method to assign Affymetrix probe sets to genes, using mapping information provided by Ensembl (release 65).

Defining transcription components. We performed PCA to define a regulatory model for the mRNA transcriptome. For each of the four data sets (human$_{\text{small}}$, human$_{\text{large}}$, rat and mouse), we computed an eigen decomposition of the Pearson product-moment correlation between all *n* probe sets over all *m* samples. This resulted in *n* eigenvalues and *n* corresponding eigenvectors (hereafter called transcriptional components, TCs). Each TC captures a part of the variance seen in the correlation space of *n* probe sets. We hypothesized that each TC is associated with an underlying factor regulating mRNA expression (and a corresponding cellular state). Each TC is composed of *n* scalars called 'coefficients' representing the relative weights of the probe sets in the TC. The inner product of a TC and the per–probe set $N(0,1)$ standardized expression vector of an individual sample gives the TC score. A TC score can be seen as an indirect measurement of the activity of the underlying regulatory factor in the sample under investigation. The sign of a probe set coefficient in a TC defines the direction of the change in mRNA expression in relation to the TC score. To determine the internal consistency and, therefore, the reliability of the identified TCs, we calculated a Cronbach's α value and split-half reliability for each TC. Pairwise Spearman's rank correlations were calculated between TCs from the different platforms to estimate the concordance between the independently defined TCs. To validate that the identified TCs were not merely mathematical constructs but were biologically relevant and to gain more biological insight into the identified TCs and their associated underlying regulatory factors, we performed GSEA. In addition, we calculated the autocorrelation per TC to identify TCs that captured the downstream consequences of cytogenetic alterations (genomic deletions and duplications) on mRNA expression levels.

Gene coregulation network analysis. We calculated a gene set enrichment score (expressed as a *z* score) for every gene set in 4 different databases (GO, KEGG, Reactome and Biocarta) for each of the 2,206 TCs identified by the methods described above. These 2,206 *z* scores formed a specific '*z*-score profile' for a gene set (for example, a biological pathway). This enabled us to investigate whether, for individual genes not previously known to be part of this gene set, the 2,206 individual TC coefficients were similar to the 2,206 *z* scores, using a Pearson correlation coefficient. Strong positive correlation between the TC coefficients for an individual gene and the *z*-score profile meant that the individual gene was behaving similarly to the gene set, indicating that the gene is likely to have a role in that gene set.

FEN1 and homologous recombination. MCF-7 or HeLa cells stably transfected with the pDR-GFP homologous recombination reporter were transfected with the indicated siRNA oligonucleotides or treated with chemical FEN1 inhibitor and subsequently transfected with the I-Sce1 endonuclease. The GFP levels of viable cells were determined by flow cytometry. γ-H2AX analysis by microscopy and flow cytometry was performed to assess DNA break formation after combined FEN1 and PARP1 inhibition. Cytotoxic effects after combined FEN1 and PARP1 inhibition were measured using clonogenic survival analysis.

Functional genomic mRNA profiling. Expression data for each individual sample were reconstructed by taking the inner product of the vector of the TC probe set coefficients and the vector of the TC scores for an individual sample, after reweighting each of the TCs according to the extent of autocorrelation. These reconstructed profiles were used to quantify the number of deletions and duplications per individual sample (see the **Supplementary Note** for a detailed description). On the basis of these reconstructed profiles, a subset of 18,713 samples with the lowest number of cytogenetic aberrations out of all 37,427 samples (human$_{\text{large}}$) was selected. This subset served as the input for a new regulatory model containing TCs that captured the effect of physiological factors on mRNA expression. A total of 718 of these non-cytogenetic TCs were used as covariates in multiple linear regression to correct the original expression profiles. The corrected profiles were called FGM profiles. In these profiles, the variance in mRNA expression due to physiological factors was minimized, emphasizing variance in expression due to genetic aberrations.

Identification of 16,172 unrelated, patient-derived tumor samples. FGM profiling is particularly attractive because of the public availability of thousands of cancer samples whose expression has been profiled with mRNA microarrays. These cancer samples harbor many genetic aberrations due to genomic instability. We decided to construct a large data set to define the FGM landscape in cancer. First, we had to ensure that we only investigated cancer samples that were not genetically identical and that did not represent cell line samples. To do so, we developed a new method to remove cell line samples, as these samples might not reflect the *in vivo* context of cancer cells. Next, we removed duplicate individuals: as is common for genome-wide association analysis, duplicate samples can lead to false positive associations and should be removed. Finally, we confined our analysis to only the samples that showed some genomic instability, as these were likely to be cancer samples. Each of these steps required the development of new methods to make this possible. A detailed description of these new methods is presented in the **Supplementary Note**.

Genome-wide gene association with genomic instability. We performed a genome-wide association analysis between individual genes (FGM expression signals) and the degree of genomic instability in 15,204 unrelated, patient-derived tumor samples. Association was determined by the Pearson product-moment correlation coefficient within meta-analysis batches. Meta-analysis *P* values were calculated according to the Liptak trend method (*z*-transformed *P* values, weighted according to the square root of the number of samples in a meta-analysis batch).

Predicting gene amplifications and deletions. For each individual gene across 41 tumor types, we quantified the percentage of samples with a significantly increased or decreased signal. To determine the thresholds that defined such signals, we applied our method to the subset of 18,713 non-cancer samples in the human$_{\text{large}}$ data and calculated the 2.5th and 97.5th percentiles for every individual gene. For each of the 16,172 tumor samples, genes were marked as significantly suppressed or amplified when the signal was below the 2.5th or above the 97.5th percentile, respectively, in non-cancer samples.

32. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
33. Crijns, A.P.G. *et al.* Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS Med.* **6**, e24 (2009).
34. Heijink, D.M. *et al.* A bioinformatical and functional approach to identify novel strategies for chemoprevention of colorectal cancer. *Oncogene* **30**, 2026–2036 (2011).

3

HUMAN DISEASE-ASSOCIATED GENETIC VARIATION IMPACTS LARGE INTERGENIC NON-CODING RNA EXPRESSION

PLoS Genetics, 9(1):e1003201, Jan 2013

Vinod Kumar, Harm-Jan Westra, Juha Karjalainen, Daria V. Zhernakova, Tõnu Esko, Barbara Hrdlickova, Rodrigo Almeida, Alexandra Zhernakova, Eva Reinmaa, Urmo Võsa, Marten H. Hofker, Rudolf S. N. Fehrmann, Jingyuan Fu, Sebo Withoff, Andres Metspalu, Lude Franke & Cisca Wijmenga

Supplementary information is available at https://goo.gl/4CC7AM

Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression

Vinod Kumar[1], Harm-Jan Westra[1]◔, Juha Karjalainen[1]◔, Daria V. Zhernakova[1]◔, Tõnu Esko[2], Barbara Hrdlickova[1], Rodrigo Almeida[1,3], Alexandra Zhernakova[1], Eva Reinmaa[2], Urmo Võsa[2], Marten H. Hofker[4], Rudolf S. N. Fehrmann[1], Jingyuan Fu[1], Sebo Withoff[1], Andres Metspalu[2], Lude Franke[1¶], Cisca Wijmenga[1¶*]

1 Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands, 2 Institute of Molecular and Cell Biology and Estonian Genome Center, University of Tartu, Tartu, Estonia, 3 Graduate Program in Health Sciences, University of Brasilia School of Health Sciences, Brasilia, Brazil, 4 Molecular Genetics Section, Department of Pathology and Medical Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Abstract

Recently it has become clear that only a small percentage (7%) of disease-associated single nucleotide polymorphisms (SNPs) are located in protein-coding regions, while the remaining 93% are located in gene regulatory regions or in intergenic regions. Thus, the understanding of how genetic variations control the expression of non-coding RNAs (in a tissue-dependent manner) has far-reaching implications. We tested the association of SNPs with expression levels (eQTLs) of large intergenic non-coding RNAs (lincRNAs), using genome-wide gene expression and genotype data from five different tissues. We identified 112 *cis*-regulated lincRNAs, of which 45% could be replicated in an independent dataset. We observed that 75% of the SNPs affecting lincRNA expression (lincRNA *cis*-eQTLs) were specific to lincRNA alone and did not affect the expression of neighboring protein-coding genes. We show that this specific genotype-lincRNA expression correlation is tissue-dependent and that many of these lincRNA *cis*-eQTL SNPs are also associated with complex traits and diseases.

Citation: Kumar V, Westra H-J, Karjalainen J, Zhernakova DV, Esko T, et al. (2013) Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. PLoS Genet 9(1): e1003201. doi:10.1371/journal.pgen.1003201

Editor: Vivian G. Cheung, University of Pennsylvania, United States of America

Received August 13, 2012; **Accepted** November 13, 2012; **Published** January 17, 2013

Funding: This study was supported by grants from the Celiac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government (grant BSIK03009 to CW)), the Netherlands Organization for Scientific Research (NWO-VICI grant 918.66.620 to CW, NWO-VENI grant 916.10.135 to LF), the Dutch Digestive Disease Foundation (MLDS WO11-30 to CW), and a Horizon Breakthrough grant from the Netherlands Genomics Initiative (grant 92519031 to LF). EGCUT received financing from FP7 programs (ENGAGE, OPENGENE), targeted financing from Estonian Government SF0180142s08, Estonian Research Roadmap through Estonian Ministry of Education and Research, Center of Excellence in Genomics (EXCEGEN), and University of Tartu (SP1GVARENG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: c.wijmenga@umcg.nl

◔ These authors contributed equally to this work.

¶ These authors were joint senior authors on this work.

Introduction

It is now evident that most of the human genome is transcribed to produce not only protein-coding transcripts but also large numbers of non-coding RNAs (ncRNAs) of different size [1,2]. Well-characterized short ncRNAs include microRNAs, small interfering RNAs, and piwi-interacting RNAs, whereas the large intergenic non-coding RNAs (lincRNAs) make up most of the long ncRNAs. LincRNAs are non-coding transcripts of more than 200 nucleotides long; they have an exon-intron-exon structure, similar to protein-coding genes, but do not encompass open-reading frames [3]. The recent description of more than 8,000 lincRNAs makes these the largest subclass of the non-coding transcriptome in humans [4].

Evidence is mounting that lincRNAs participate in a wide-range of biological processes such as regulation of epigenetic signatures and gene expression [5–7], maintenance of pluripotency and differentiation of embryonic stem cells [8]. In addition, several individual lincRNAs have also been implicated in human diseases. A well-known example is a region on chromosome 9p21 that encompasses an antisense lincRNA, ANRIL (antisense lincRNA of the *INK4* locus). Genome-wide association studies (GWAS) have shown that this region is significantly associated with susceptibility to type 2 diabetes, coronary disease, and intracranial aneurysm as well as different types of cancers [9] and some of the associated SNPs have been shown to alter the transcription and processing of ANRIL transcripts [10]. Similarly, increased expression of lincRNA HOTAIR (HOX antisense non-coding RNA) in breast cancer is associated with poor prognosis and tumor metastasis [10]. Another example is MALAT-1 (metastasis associated in lung adenocarcinoma transcript) where the expression is three-fold higher in metastasizing tumors of non-small-cell lung cancer than in non-metastasizing tumors [11].

In addition, over the last decade, more than 1,200 GWAS have identified nearly 6,500 disease- or trait-predisposing SNPs, but only 7% of these are located in protein-coding regions [12,13]. The remaining 93% are located within non-coding regions [14], suggesting that GWAS-associated SNPs regulate gene transcription levels rather than altering the protein-coding sequence or protein structure. Even though there is growing evidence to

Author Summary

Large intergenic non-coding RNAs (lincRNAs) are the largest class of non-coding RNA molecules in the human genome. Many genome-wide association studies (GWAS) have mapped disease-associated genetic variants (SNPs) to, or in, the vicinity of such lincRNA regions. However, it is not clear how these SNPs can affect the disease. We tested whether SNPs were also associated with the lincRNA expression levels in five different human primary tissues. We observed that there is a strong genotype-lincRNA expression correlation that is tissue-dependent. Many of the observed lincRNA *cis*-eQTLs are disease- or trait-associated SNPs. Our results suggest that lincRNA-eQTLs represent a novel link between non-coding SNPs and the expression of protein-coding genes, which can be exploited to understand the process of gene-regulation through lincRNAs in more detail.

implicate lincRNAs in human diseases [15,16], it is unknown whether disease-associated SNPs could affect the expression of non-coding RNAs. We hypothesized that GWAS-associated SNPs can affect the expression of lincRNA genes, thereby proposing a novel disease mechanism.

To test this hypothesis, we performed eQTL mapping on 2,140 human lincRNA-probes using genome-wide gene expression and genotype data of 1,240 peripheral blood samples (discovery cohort) [17]. The lincRNA *cis*-eQTLs identified were then tested for replication in an independent cohort containing 891 peripheral blood samples (replication cohort). Since lincRNAs are considered to be more tissue-specific than protein-coding genes [4], we set-out to identify tissue-dependent *cis*-eQTLs for lincRNAs using data from another four different primary tissues from the subset of 85 individuals in our primary cohort [18]. Subsequently, we tested whether SNPs that affect the levels of lincRNA expression are associated with diseases or traits. Finally, we predicted the most likely function(s) of a subset of *cis*-eQTL lincRNAs by using co-regulation information from a compendium of approximately 80,000 expression arrays (www.GeneNetwork.nl).

Results

Commercial microarrays contain probes for a subset of non-coding RNA

Whole-genome gene expression oligonucleotide arrays have played a crucial role in our understanding of gene regulatory networks. Even though most of the currently available commercial microarrays are designed to capture all known protein-coding transcripts, they still include subsets of probes that capture transcripts of unknown function (sometimes abbreviated as TUF). We investigated whether the TUF probes present on the Illumina Human HT12v3 array, overlap with lincRNA transcripts that were recently described in the lincRNA catalog [4]. The lincRNA catalog contained a provisional set of 14,393 transcripts mapping to 8,273 lincRNA genes and a stringent set of 9,918 transcripts mapping to 4,283 lincRNA genes. We identified 2,140 unique probes that map to 1,771 different lincRNAs from the provisional set and 1,325 unique probes that map to 1,051 lincRNA genes from the stringent set. We chose 2,140 unique probes that mapped to lincRNAs from the provisional set for further eQTL analysis.

Genetic control of lincRNAs expression in blood

It is known that in general lincRNAs are less abundantly expressed compared to protein-coding transcripts [4]. To test the expression levels of the 2,140 lincRNA probes in 1,240 peripheral blood samples (discovery cohort), we compared the quantile-normalized, log scale transformed mean expression intensity as well as expression variation of the lincRNA probes to probes mapping to protein-coding transcripts. We indeed observed a significant difference in the expression levels, where lincRNA probes are less abundant (mean expression = 6.67) than probes mapping to protein-coding transcripts (mean expression = 6.92, Wilcoxon Mann Whitney $P<2.2\times10^{-16}$; Figure S1). We also observed a highly significant difference in the expression variation between lincRNA probes and probes mapping to protein-coding transcripts (Wilcoxon Mann Whitney $P<3.85\times10^{-96}$). Next, we tested whether the expression of these 2,140 lincRNA probes is affected by SNPs in *cis*, by performing eQTL mapping in these 1,240 peripheral blood samples for which genotype data was also available. We confined our analysis to SNP-probe combinations for which the distance from the center of the probe to the genomic location of the SNP was ≤250 kb. In the end, at a false-discovery rate (FDR) of 0.05, we identified 5,201 significant SNP-probe combinations, reflecting 4,644 different SNPs; these affected the expression of 112 out of 2,140 different lincRNA probes. The 112 lincRNA probes mapped to 108 lincRNA genes and comprised 5.2% of all tested lincRNA probes, with a nominal significance ranging from $P<2.8\times10^{-4}$ to 9.81×10^{-198} in peripheral blood (Table S1).

Replication of lincRNA *cis*-eQTLs in an independent blood dataset

We then performed a replication analysis to test the reproducibility of the identified 112 lincRNA *cis*-eQTLs using an independent dataset of 891 whole peripheral blood samples. We took the 112 lincRNA-probes (or 5,201 SNP-probe pairs) that were significantly affected by *cis*-eQTLs in the discovery cohort and tested whether these eQTLs were also significant in the replication dataset (at FDR 0.05). We could replicate 45% of the 112 lincRNA *cis*-eQTLs at an FDR<0.05, of which all the eQTLs had an identical allelic direction (Figure S2). The smaller sample size of the replication cohort compared to the discovery cohort makes it inherently difficult to replicate all the *cis*-eQTLs that we have detected in the discovery cohort.

Number of *cis*-eQTLs is dependent on expression levels of transcripts

Our observation that 5.2% of all tested lincRNAs are *cis*-regulated (Table S1) might seem disappointing, compared to our earlier observation that 25% of the protein-coding probes in this dataset are *cis*-regulated [18]. However, we reasoned that the generally lower expression levels of lincRNAs compared to protein-coding genes might make it more difficult to detect *cis*-eQTLs for lincRNAs, as the influence of background noise becomes substantial for less abundant transcripts, making accurate expression quantification difficult (Figure S1A).

Indeed, we found significantly higher expression levels for the 112 *cis*-eQTL lincRNA probes (mean expression = 6.80) compared to the 2,028 non-eQTL lincRNA probes (mean expression = 6.66 Wilcoxon Mann Whitney $P=3.88\times10^{-15}$; Figure S3) and also observed a significant difference in expression variance between the 112 *cis*-eQTL lincRNAs compared to the 2,028 non-*cis* eQTL lincRNAs (Wilcoxon Mann Whitney $P=1.067\times10^{-8}$), indicating that lower overall expression levels do make identification of *cis*-eQTLs more difficult.

To further confirm the relationship between average expression levels of probes and the number of detectable *cis*-eQTLs, we first

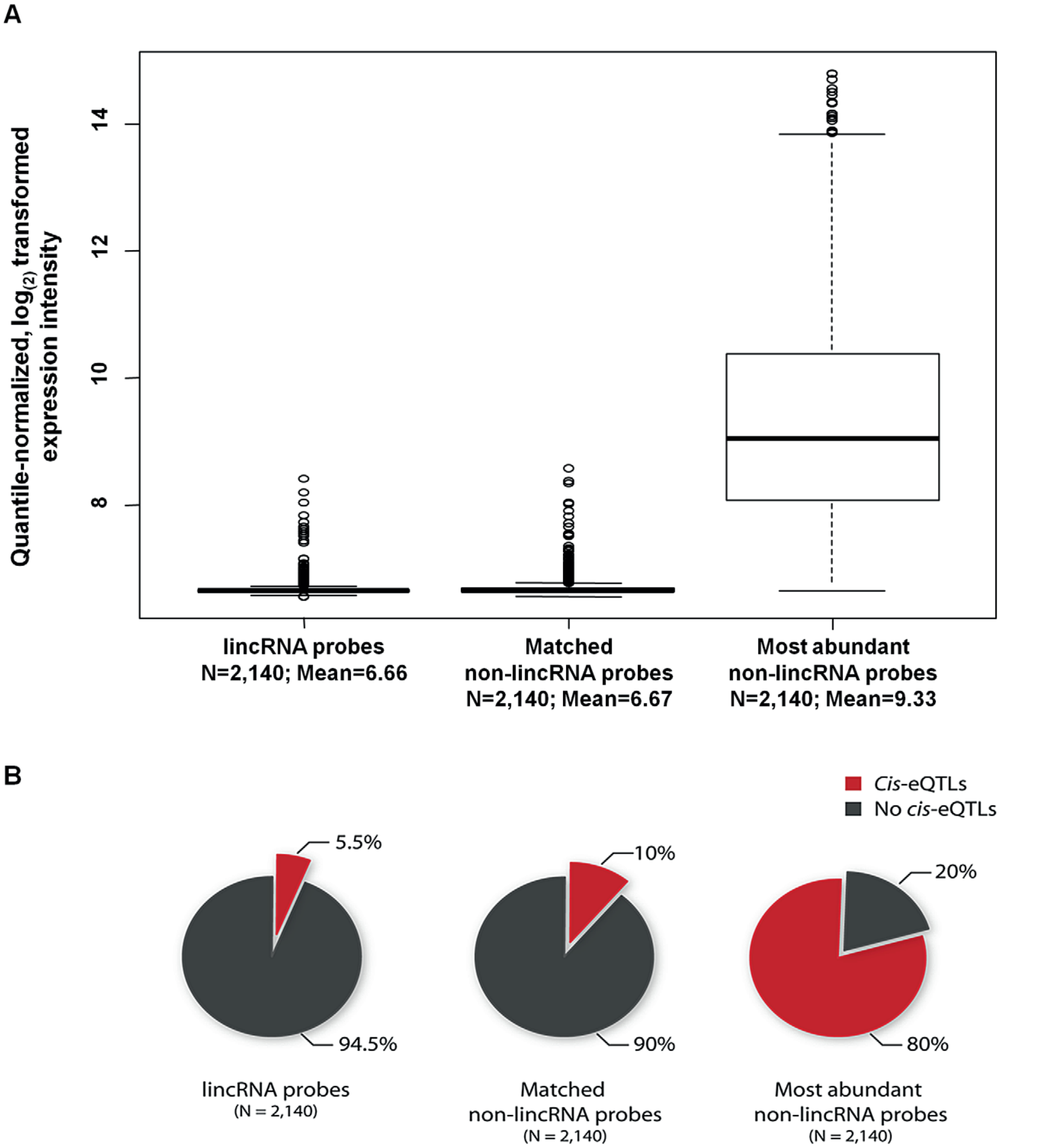


Figure 1. The number of detected *cis*-eQTLs is dependent on the expression levels of the transcripts. (A) Quantile-normalized average expression intensity and (B) number of *cis*-eQTL affected probes in percentage, for 2,140 lincRNA probes, 2,140 non-lincRNA (matched for 2,140 lincRNA probes' median expression and standard deviation) and 2,140 most abundantly expressed non-lincRNA probes.
doi:10.1371/journal.pgen.1003201.g001

mapped *cis*-eQTLs for an equal set of 2,140 probes that were instead protein-coding and were the most abundantly expressed of all protein-coding probes. We also conducted *cis*-eQTL mapping for a set of 2,140 protein-coding probes that had been selected to have an identical expression intensity distribution as the 2,140 lincRNA probes (i.e. matched for mean expression intensity and standard deviation), using the same 1,240 blood samples (Figure 1A). We indeed observed a profound relationship between average expression levels of protein-coding transcripts and the number of detectable *cis*-eQTLs. Eighty percent of the 2,140 most abundantly expressed protein-coding probes showed a *cis*-eQTL effect, whereas only 10% of the protein-coding probes that had

Table 1. Some of the lincRNA *cis*-eQTLs are disease-associated SNPs.

Cis-eQTL SNP	eQTL *P* on lincRNA	Proxies (R^2>0.8) associated with disease/trait	Chr	Trait/Disease	eQTL affected lincRNA	eQTL tissue
rs13278062	4.31×10^{-32}	rs13278062	8	Exudative age-related macular degeneration	XLOC_006742	Blood
rs11066054	4.09×10^{-11}	rs6490294	12	Mean platelet volume	XLOC_010202	Blood
rs206942	3.63×10^{-5}	rs206936	6	Body mass index	XLOC_005690	Blood
rs11065766	6.67×10^{-5}	rs10849915	12	Alcohol consumption	XLOC_009878	Blood
	6.67×10^{-5}	rs10774610	2	Drinking behavior		
rs1465541	1.84×10^{-4}	rs11684202	2	Coronary heart disease	XLOC_002026	Blood
rs12125055	1.84×10^{-4}	rs7542900	1	Type 2 diabetes	XLOC_000922	Blood
rs199439	8.25×10^{-6}	rs199515	17	Parkinson's disease	XLOC_012496	SAT
		rs415430	17	Parkinson's disease		SAT
		rs199533	17	Parkinson's disease		SAT
rs17767419	1.05×10^{-8}	rs17767419	16	Thyroid volume	XLOC_011797	SAT, VAT
		rs3813582	16	Thyroid function		SAT, VAT

Chr chromosome, SAT Saturated adipose tissue, VAT Visceral adipose tissue.

doi:10.1371/journal.pgen.1003201.t001

been matched for an expression intensity of the 2,140 lincRNA-probes were affected by *cis*-eQTLs (Figure 1B).

Hence it is possible that if we can accurately quantify all lincRNAs in large RNA-sequencing datasets, we will be able to identify *cis*-eQTLs for a larger proportion of all lincRNAs.

Most SNPs that affect lincRNA expression do not alter the expression of protein-coding genes

It could be possible that the SNPs that affect lincRNA expression actually operate by first affecting protein-coding gene expression levels, which in turn affect lincRNA expression. If this were to be the case, our identified lincRNA *cis*-eQTLs would merely be a by-product of protein-coding *cis*-eQTLs. To ascertain this, we tested whether the 112 lincRNA-eQTL SNPs were also significantly affecting neighboring protein-coding genes. By keeping the same significance threshold (at FDR<0.05 level, the *P*-value threshold was 2.4×10^{-4}), we observed that nearly 75% (83 out of 112) of the lincRNA-eQTLs were affecting only lincRNAs, even though the interrogated neighboring protein-coding genes were generally more abundantly expressed than the lincRNAs themselves (Figure S4). Genetic variants can thus directly regulate the expression levels of lincRNAs.

We found 29 *cis*-eQTLs to be associated with the expression of both lincRNA and protein coding genes. For 50% of these 29 *cis*-eQTLs, we found that the expression of lincRNAs and protein-coding genes was in the opposite direction, whereas for the other 50% of *cis*-eQTLs, both types of transcripts were co-regulated in the same direction (Figure S5).We tested whether these 29 *cis*-eQTLs are the strongest eQTLs for both lincRNA and protein-coding genes. Although these 29 *cis*-eQTLs were the strongest eQTLs for lincRNAs, only 5 among 29 were also the strongest eQTLs for protein-coding genes. This observation further highlights the direct regulation of lincRNA expression through genetic variants.

Some lincRNA *cis*-eQTLs are tissue-dependent

There is considerable interest in mapping eQTLs in disease-relevant tissue types. We reasoned that since expression of the lincRNAs seems to be much more tissue-specific than the expression of protein-coding genes [4], mapping lincRNA-eQTLs in different tissues could reveal additional, tissue-specific lincRNA-eQTLs. To test this, we analyzed gene expression and genotype data of 74 liver samples, 62 muscle samples, 83 subcutaneous adipose tissue (SAT) samples, and 77 visceral adipose tissue (VAT) samples from our primary cohort of 85 unrelated, obese Dutch individuals [18]. Upon *cis*-eQTL mapping we detected 35 *cis*-eQTL-probes, of which 18 were specific in the four different non-blood tissues, resulting in a total of 130 lincRNA-eQTLs in the combined set of all five tissues (Table S1). Five *cis*-eQTLs identified in blood tissue were also significantly replicated in at least one other non-blood tissue (Table S1). While we could replicate 45% of the *cis*-eQTLs in the substantial whole peripheral blood replication cohort, the replication rate in the very small cohorts for fat, liver and muscle tissue was, as expected, much lower. We were able to observe tissue-specific lincRNA eQTLs in muscle (1), liver (4), SAT (9) and blood (107) (Figure S6). Since the four non-blood tissue expression levels were from the same individuals, these results do indeed provide evidence that some of the lincRNAs are regulated by genetic variants in a tissue-specific manner.

LincRNA tissue specific *cis*-eQTLs are disease-associated SNPs

As most of the GWAS-associated SNPs are located within non-coding regions, we tested whether the 130 lincRNA-eQTLs identified in five different tissues are also GWAS-associated variants. To do this, we intersected trait-associated SNPs (at reported nominal $P<9.9\times10^{-6}$, retrieved from the catalog of published genome-wide association studies per 26 July 2012) [14] with the 130 top lincRNA *cis*-eQTLs and their proxies (proxies with $R^2>0.8$ using the 1000Genome CEU population as reference). We identified 12 GWAS SNPs or their proxies, that were also a lincRNA *cis*-eQTLs of eight different lincRNA genes (Table 1). All except one of the 12 SNPs were exclusively associated with lincRNA expression and thus did not affect the expression levels of neighboring protein-coding genes (Table 1), suggesting a causative role of altered lincRNA expression for these phenotypes.

Notably SNP rs13278062 at 8p21.1, associated with exudative age-related macular degeneration (AMD) in the Japanese population, was reported to alter the transcriptional levels of *TNFRSF10A* (Tumor necrosis factor receptor superfamily 10A) protein-coding gene [19]. Here we identified SNP rs13278062 as a highly significant *cis*-eQTL of lincRNA XLOC_006742 (LOC389641) ($P = 4.31 \times 10^{-32}$) rather than for *TNFRSF10A* ($P = 4.21 \times 10^{-4}$) protein-coding gene (Figure S7). Furthermore, SNP rs13278062 is located in exon 1 of lincRNA XLOC_006742, which encompasses an ENCODE (Encyclopedia of DNA elements) enhancer region characterized by H3K27acetylation and DNaseI hypersensitive clusters [20] (Figure S8).

Another interesting example is at 17q21.31 where three Parkinson's disease associated SNPs were in strong linkage disequilibrium ($R^2 > 0.8$) with top *cis*-eQTL SNP rs199439, which affects lincRNA XLOC_012496 expression exclusively in SAT (Table 1). Weight loss due to body-fat wasting is a very common but poorly understood phenomenon in Parkinson's disease patients [21]. In this regard, it is intriguing to note that the Parkinson's disease associated SNPs affects lincRNA expression exclusively in fat tissue (Table 1). Hence, identifying lincRNA-eQTLs in disease-relevant tissue types using larger groups of individuals may open up new avenues towards achieving a better understanding of disease mechanisms.

LincRNA function predictions using a co-expression network of ~80,000 arrays: A mechanistic link between disease and lincRNA

Our observations suggest a role for lincRNAs in complex diseases and other phenotypes. The next, rather daunting task is to elucidate the function of these ncRNAs. We recently developed a co-regulation network (GeneNetwork, www.genenetwork.nl/genenetwork, *manuscript in preparation*), to predict the function of any transcript based on co-expression data extracted from approximately 80,000 Affymetrix microarray experiments (see Methods). We interrogated the GeneNetwork database to predict the function of eQTL-affected lincRNAs. Among the 130 *cis*-eQTL lincRNAs that we had identified in the five different tissues, 43 were represented by expression probe sets on Affymetrix arrays for which we could predict the function (Table S2). These 43 probes include four out of eight disease-associated lincRNAs described above (Table 1) and function prediction for these probes provided relevant biological explanations.

LincRNA co-expression analysis: Disease-associated lincRNAs are co-expressed with neighboring protein-coding genes

It has been reported that some transcribed long ncRNAs function as enhancers that regulate the expression of neighboring genes [3] and may thereby contribute to the disease pathology. We found that the AMD-associated lincRNA XLOC_006742 (LOC389641) (by virtue of SNP rs13278062 which exhibits a significant eQTL effect) (Figure S7) is in strong co-expression with *TNFRSF10A* based on our GeneNetwork database (Table S3). AMD is a leading cause of blindness among elderly individuals worldwide and recent studies, both in animal models and in humans, provide compelling evidence for the role of immune system cells in its pathogenesis [22]. The gene *TNFRSF10A*, which encodes TRAIL receptor 1 (TRAIL1), has been implicated as a causative gene for AMD [19]. It has been shown that binding of TRAIL to TRAILR1 can induce apoptosis through caspase 8 activation [23] and using GeneNetwork we also predict a role in apoptosis for lincRNA XLOC_006742 (Table S2).

Another trait-associated SNP, rs11065766, is the top *cis*-eQTL of lincRNA XLOC_009878 (ENSG00000185847 or RP1-46F2.2 or LOC100131138) and it is in strong linkage disequilibrium with two SNPs associated with alcohol drinking behavior (Table 1). We found that the lincRNA XLOC_009878 is strongly co-expressed with the neighboring protein-coding gene *MYL2* (Table S4) and, according to our predictions, lincRNA XLOC_009878 is involved in striated muscle contraction ($P = 1.22 \times 10^{-26}$). Chronic alcohol abuse can lead to striking changes in skeletal muscle structure, which in turn plays a role in the development of alcoholic myopathy and/or cardiomyopathy [24]. It has also been reported that alcohol can reduce the content of skeletal muscle proteins such as titin and nebulin to affect muscle function in rats [25]. We found lincRNA XLOC_009878 to be co-expressed with titin and many other skeletal muscle proteins necessary for the structural integrity of the muscle (Table S4). Thus, it needs to be tested whether deregulation of lincRNA XLOC_009878 expression might alter an individual's ability to metabolize alcohol due to changes in the muscle functional property.

Localization of lincRNA *cis*-eQTLs in regulatory regions

We found that more than 70% of the lincRNA *cis*-eQTLs from both blood and non-blood tissues were located in intergenic regions with respect to protein-coding genes (Figure 2A). We also found high frequencies of lincRNA *cis*-eQTLs to be located around transcriptional start site (Figure 2B), suggesting that these *cis*-eQTLs may affect the expression of lincRNAs through similar gene regulatory mechanisms as those seen for protein-coding *cis*-eQTLs. Thus, in order to understand the mechanism of how lincRNA *cis*-eQTLs affect lincRNA expression, we intersected the location of top 112 lincRNA *cis*-eQTLs and their proxies ($r^2 = 1$) in blood with regulatory regions using the HaploReg database [26]. The results suggested that indeed most of the lincRNA *cis*-eQTLs (69%) were located in functionally important regulatory regions (Figure S8), which contained DNAse I regions, transcription factor binding regions, and histone marks of promoter and enhancer regions. Furthermore, these *cis*-eQTLs were found to be located more often within blood cell-specific enhancers (K562 and GM12878) (Figure 3A), suggesting that some of these *cis*-eQTLs regulate lincRNA expression in a tissue-specific manner through altering these enhancer sequences.

Since we observed enrichment of cell-specific enhancers for lincRNA *cis*-eQTLs within blood cells (K562 and GM12878), we compared the fold enrichment of enhancers in these two cell types to see whether lincRNA *cis*-eQTLs are more often located in functionally important regions than any random set of SNPs. We found a significant difference in the enrichment of enhancers in which more than a 4-fold enrichment was seen for real *cis*-eQTLs both in K562 cells ($P = 0.0004$) and GM12878 cells ($P = 0.011$) compared to permuted SNPs. These findings suggest that some of the identified lincRNA *cis*-eQTLs are indeed functional SNPs.

Discussion

Even though it may have been expected that lincRNA expression would be under genetic control, this is the first study, to our knowledge, to comprehensively establish this link. We were able to identify *cis*-eQTLs in five different tissues and have demonstrated that common genetic variants regulate the expression of lincRNAs alone. It is intriguing that around 75% of lincRNA *cis*-eQTLs are specific to lincRNAs alone, but not to protein-coding genes. Recent data from the ENCODE project suggests that combinations of different transcription factors are involved in regulating gene-expression in different cell types and

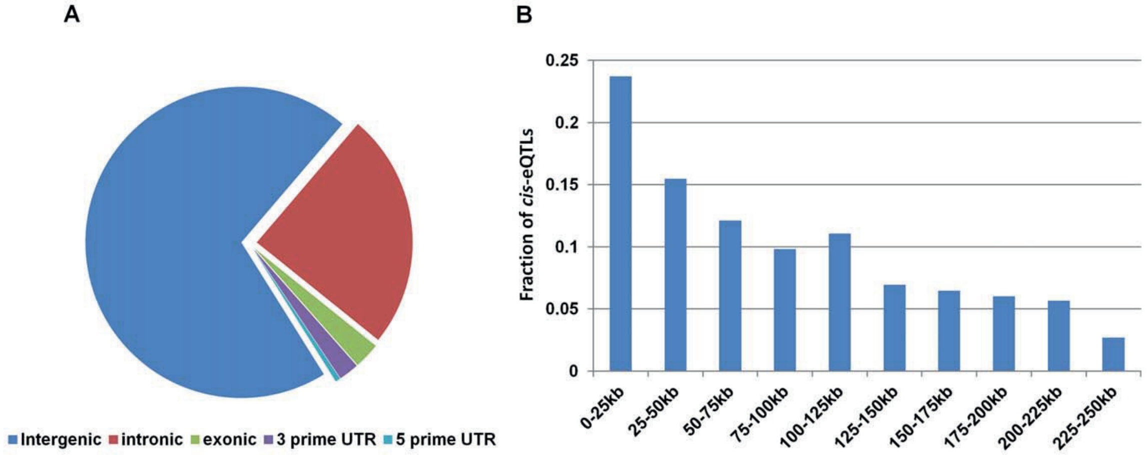


Figure 2. Distribution of lincRNA *cis*-eQTLs with respect to different transcripts. (A) The majority of the lincRNA *cis*-eQTLs are located within the non-coding part of the genome and less than 6% of lincRNA *cis*-eQTLs are located within mRNA. (B) Distribution of lincRNA *cis*-eQTLs with respect to distance to the lincRNA transcripts. The x-axis displays the 250 kb window used for *cis*-eQTL mapping and the y-axis displays the fraction of lincRNA *cis*-eQTLs located within this window.
doi:10.1371/journal.pgen.1003201.g002

non-coding RNAs tend to be regulated by certain combinations of transcription factors more often than others [27]. Thus, it could still be possible that some transcription factors specifically regulate lincRNA expression. We also observed a strong relationship between whether or not a transcript is affected by *cis*-eQTLs and its expression levels, where highly abundant transcripts were more often affected by *cis*-eQTLs. This relationship was comparable between lincRNA and protein-coding probes, although protein-coding probes (matched for expression levels of lincRNA probes) tend to show more *cis*-eQTLs (Figure 1B; 5.2% versus 10%). Although this difference is not drastic, it may suggest that lincRNAs exhibit another layer of gene regulation which is more tissue-specific. Thus, we may expect to identify many more lincRNA *cis*-eQTLs once larger datasets of different tissues become available.

One limitation of our study is the lack of probes to comprehensively map eQTLs to all the reported lincRNAs, as we relied upon microarrays. Future analyses using RNA-sequencing datasets will undoubtedly provide much more insight into how genetic variants affect lincRNA expression. So far, two landmark RNA-sequencing based eQTL studies have been published using 60 (Montgomery et al) [28] and 69 samples (Pickrell et al) [29], respectively. While Pickrell et al did not mention lincRNAs with a *cis*-eQTL effect, Montgomery et al identified six *cis*-regulated

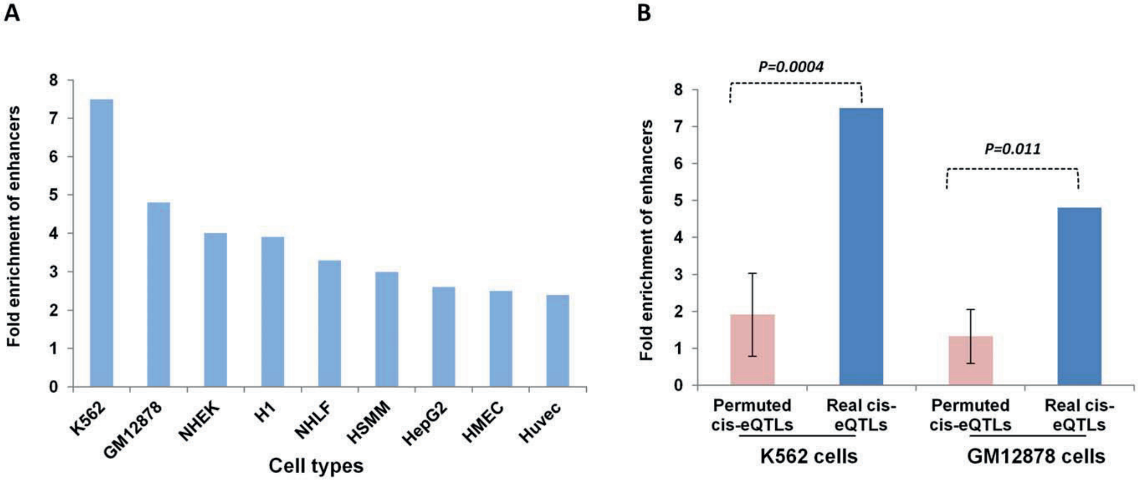


Figure 3. Localization of lincRNA *cis*-eQTLs in regulatory regions. (A) A plot to indicate the location of lincRNA *cis*-eQTLs in cell-specific enhancers. The x-axis shows the different cell lines analyzed and the y-axis shows the fold enrichment of enhancers. (B) A plot to show the difference in fold enrichment of enhancers for real lincRNA *cis*-eQTLs compared to permuted lincRNA *cis*-eQTLs. The significance of the difference in fold enrichment was tested by T-test. The HaploReg database was used to analyze the fold enrichment of enhancers.
doi:10.1371/journal.pgen.1003201.g003

lincRNAs (at a slightly higher FDR of 0.17). We re-analyzed these two datasets and found that we could replicate one of the 112 *cis*-eQTL lincRNAs effects that we detected using arrays (with an identical allelic direction; Figure S10). These results indicate that *cis*-eQTL lincRNAs detected using conventional microarrays can be replicated in sequencing-based datasets. However, it also indicates that sample size is currently a limiting factor in finding many more *cis*-eQTL lincRNAs in sequencing-based datasets.

Nevertheless, our results clearly indicate that there is a strong genotype-lincRNA expression correlation that is tissue-dependent. A considerable number of the observed lincRNA *cis*-eQTLs are disease- or trait-associated SNPs. Since lincRNAs can regulate the expression of protein-coding genes either in *cis* [3] or in *trans* [8], lincRNA-eQTLs represent a novel link between non-coding SNPs and the expression of protein-coding genes. Our examples show that this link can be exploited to understand the process of gene-regulation in more detail, which may assist us in characterizing lincRNAs as another class of disease biomarkers.

Methods

Ethics statement

This study was approved by the Medical Ethical Board of Maastricht University Medical Center (four non-blood tissues), and local ethical review boards (1,240 peripheral blood samples) in line with the guidelines of the 1975 Declaration of Helsinki. Informed consent in writing was obtained from each subject personally. The subject information is provided in Table S5.

Mapping probes to lincRNAs

A detailed mapping strategy of Illumina expression probe sequences has been described previously [17]. We extracted 43,202 expression probes mapping to single genomic locations (hg18 build) and excluded those that did not map or that mapped to multiple different loci. LincRNA chromosomal coordinates (hg19 build) were obtained from the lincRNA catalog (http://www.broadinstitute.org/genome_bio/human_lincrnas/?q = lincRNA_catalog) and converted to hg18 coordinates using UCSC's LiftOver application (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Subsequently, we extracted probes mapping to lincRNA exonic regions by employing BEDtools [30].

Blood dataset of 1,240 samples

The blood dataset and a detailed eQTL mapping strategy have been described previously [17]. Briefly, 1,240 peripheral blood samples from unrelated, Dutch control subjects were investigated (Table S5). Genotyping of these samples was performed according to Illumina's standard protocols (Illumina, San Diego, USA), using either the HumanHap370 or 610-Quad platforms. Because the non-blood samples (see below) were genotyped using Illumina HumanOmni1-Quad BeadChips, we applied IMPUTE v2 [31] to impute the genotypes of SNPs that were covered by the Omni1-Quad chip but that were not included on the Hap370 or 610-Quad platforms [31]. Anti-sense RNA was synthesized using the Ambion Illumina TotalPrep Amplification Kit (Ambion, New York, USA) following the manufacturer's protocol. Genome-wide gene expression data was obtained by hybridizing complementary RNA to Illumina's HumanHT-12v3 array and subsequently scanning these chips on the Illumina BeadArray Reader.

Replication blood dataset of 891 samples

We used a dataset comprising peripheral blood samples of 891 unrelated individuals from the Estonian Genome Centre, University of Tartu (EGCUT) biobank cohort of 53,000 samples for replication. Genotyping of these samples was performed according to Illumina's standard protocols, using Illumina Human370CNV arrays (Illumina Inc., San Diego, US), and imputed using IMPUTE v2 [31], using the HapMap CEU phase 2 genotypes (release #24, build 36). Whole peripheral blood RNA samples were collected using Tempus Blood RNA Tubes (Life Technologies, NY, USA), and RNA was extracted using Tempus Spin RNA Isolation Kit (Life Technologies, NY, USA). Quality was measured by NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, DE, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). Whole-Genome gene-expression levels were obtained by Illumina Human HT12v3 arrays (Illumina Inc, San Diego, US) according to manufacturers' protocols.

Four non-blood primary tissues

Previously we described tissue-dependent eQTLs in 74 liver samples, 62 muscle samples, 83 SAT samples and 77 VAT samples from a cohort of 85 unrelated, obese Dutch individuals (all four tissues were available for 48 individuals) [18] (Table S5). These samples were genotyped according to standard protocols from Illumina, using Illumina HumanOmni-Quad BeadChips (Omni1). Genome-wide gene expression data of all samples was assayed by hybridizing complementary RNA to the Illumina HumanHT-12v3 array and then scanning it on the BeadArray Reader.

Cis-eQTL mapping

The method for normalization and principal component analysis-based correction of expression data, along with the methods to control population stratification and SNP quality, were described previously [17,18]. The *cis*-eQTL analysis was performed on probe-SNP combinations for which the distance from the center of the probe to the genomic location of the SNP was ≤250 kb. Associations were tested by non-parametric Spearman's rank correlation test and the *P* values were corrected for multiple testing by false-discovery rate (FDR) at $P<0.05$, in which the distribution was obtained from permuting expression phenotypes relative to genotypes 100 times within the HT12v3 dataset and comparing those with the observed *P*-value distribution. At FDR = 0.05 level, the *P*-value threshold was 2.4×10^{-4} for significantly associated probe-SNP pairs in blood, 1.5×10^{-5} in SAT, 5.21×10^{-6} in VAT, 6.3×10^{-6} in liver and 1.8×10^{-6} in muscle.

LincRNA function prediction

To predict the function(s) for lincRNAs, we interrogated the GeneNetwork database (www.genenetwork.nl/genenetwork) that has been developed in our lab (*manuscript in preparation*). In short, this database contains data extracted from approximately 80,000 microarray experiments that is publically available from the Gene Expression Omnibus; after extensive quality control, it contains data on 54,736 human, 17,081 mouse and 6,023 rat Affymetrix array experiments. Principal component analysis was performed on probe-set correlation matrices of each of four platforms (two human platforms, one mouse and one rat platform), resulting in 777, 377, 677 and 375 robust principal components, respectively. Jointly these components explain between 79% and 90% of the variance in the data, depending on the species or platform. Many of these components are well conserved across species and enriched for known biological phenomena. Because of this, we were able to combine the results into a multi-species gene network with 19,997 unique human genes, allowing us to utilize the principal components to accurately predict gene function by using a 'guilt-by-association' procedure (a description of the method is available at www.genenetwork.nl/genenetwork). Predictions were made based on pathways and gene sets from Gene Ontology, KEGG, BioCarta, TransFac and Reactome.

Functional annotation of lincRNA *cis*-eQTLs

We employed the HaploReg web tool [26] to intersect SNPs (and their perfect proxies, $r^2 = 1$ using the CEU samples from the 1000 Genomes project) with regulatory information and also to calculate the fold enrichment of cell-type specific enhancers. In order to ascertain whether this enrichment was higher than expected, we took eQTL results from 100 permutations (shuffling the gene expression identifier labels): for each permutation we determined the top 112 eQTL probes and took the corresponding top SNPs and their perfect proxies ($r^2 = 1$). We extracted the fold enrichment of enhancers from HaploReg for these 100 sets of SNPs as well, which then permitted us to estimate the significance of enrichment of the real eQTL analysis, determined by fitting a normal distribution on the 100 log-transformed permutation enrichment scores.

Supporting Information

Figure S1 LincRNA probes show different expression characteristics compared to other transcripts. The figure shows the difference in quantile-normalized average expression intensity between lincRNA probes and non-lincRNA probes. The significance of difference in expression intensity was tested by the Wilcoxon Mann Whitney test.
(TIF)

Figure S2 Replicated lincRNA cis-eQTLs show identical allelic direction of effect in the both the discovery and replication datasets. We compared the z-scores (association strength) of each significantly associated probe-SNP pair in the discovery dataset (Groningen HT12v3; N = 1,240) with the replication dataset (EGCUT; N = 891).
(TIF)

Figure S3 lincRNA probes with *cis*-eQTL effect show higher expression levels compared to lincRNA probes without *cis*-eQTL effect. The significance of difference in expression intensity was tested by the Wilcoxon Mann Whitney test.
(TIF)

Figure S4 LincRNA *cis*-eQTL SNPs mostly affect lincRNA transcripts alone. Quantile-normalized average expression intensity of *cis*-eQTL lincRNAs and their neighboring protein coding genes without *cis*-eQTL.
(TIF)

Figure S5 Distribution of Z-scores of co-regulated lincRNA and protein-coding genes. We compared the z-scores (association strength) of each significantly associated probe-SNP pair for the 29 *cis*-eQTLs that affect both lincRNAs and protein-coding genes.
(TIF)

Figure S6 Number of specific and overlapping *cis*-eQTL lincRNAs identified across five different tissues.
(TIF)

Figure S7 Plots to show the association of age-related macular degeneration SNP rs13278062 with expression levels of lincRNA LOC389641 and protein-coding gene *TNFRSF10A* in blood (N = 1,249). The x-axis shows the number of samples according to the genotypes at rs13278062 and the y-axis is the average expression intensity of probes.
(TIF)

Figure S8 UCSC genome browser screen shot (*http://genome.ucsc.edu*) to show the location of age-related macular degeneration SNP, rs13278062. The x-axis is the chromosome location in the hg19 build and indicates the location of transcripts and regulatory elements identified by ENCODE on chromosome 8.
(TIF)

Figure S9 A plot to show the number of lincRNA *cis*-eQTLs on the y-axis within different regulatory regions on the x-axis.
(TIF)

Figure S10 Plots to show the *cis*-eQTL effect on lincRNA XLOC_00197 from both microarray data (Groningen HT12v3; N = 1,240) and RNA-sequencing data (Montgomery et al; N = 60). The x-axis shows the number of samples according to the genotypes at rs1120042 and rs2279692 (LD between these two SNPs, $R^2 = 0.96$) in microarray data and RNA-sequencing data, respectively.
(TIF)

Table S1 LincRNA *cis*-eQTLs in blood and four other non-blood tissues.
(XLSX)

Table S2 Function prediction of lincRNAs affected by *cis*-eQTLs using GeneNetwork.
(XLSX)

Table S3 Identification of co-expressed genes for lincRNA LOC389641 using GeneNetwork.
(XLSX)

Table S4 Identification of co-expressed genes for lincRNA LOC100131138 using GeneNetwork.
(XLSX)

Table S5 Characteristics of sample cohorts used for *cis*-eQTL mapping.
(XLSX)

Acknowledgments

The authors would like to thank Jackie Senior for editing the final text. We acknowledge EGCUT technical personnel, especially Mr. V. Soo and S. Smit. EGCUT data analyzes were carried out in part in the High Performance Computing Center of the University of Tartu.

Author Contributions

Conceived and designed the experiments: VK LF CW. Performed the experiments: VK H-JW JK DVZ TE BH RA AZ ER UV JF. Analyzed the data: VK H-JW JK DVZ. Contributed reagents/materials/analysis tools: MHH RSNF AM CW. Wrote the paper: VK H-JW SW LF CW.

References

1. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.
2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316: 1484–1488.
3. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, et al. (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143: 46–58.
4. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25: 1915–1927.

5. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A 106: 11667–11672.
6. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129: 1311–1323.
7. Nagano T, Mitchell JA, Sanz LA, Pauler FM, Ferguson-Smith AC, et al. (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. Science 322: 1717–1720.
8. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, et al. (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477: 295–300.
9. Pasmant E, Sabbagh A, Vidaud M, Bieche I (2011) ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. FASEB J 25: 444–448.
10. Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, et al. (2010) Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. PLoS Genet 6: e1001233. doi:10.1371/journal.pgen.1001233
11. Ji P, Diederichs S, Wang W, Boing S, Metzger R, et al. (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22: 8031–8041.
12. Pennisi E (2011) The Biology of Genomes. Disease risk links to gene regulation. Science 332: 1031.
13. Kumar V, Wijmenga C, Withoff S (2012) From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. Semin Immunopathol.
14. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367.
15. Martin L, Chang HY (2012) Uncovering the role of genomic "dark matter" in human disease. J Clin Invest 122: 1589–1595.
16. Jendrzejewski J, He H, Radomska HS, Li W, Tomsic J, et al. (2012) The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. Proc Natl Acad Sci U S A 109: 8646–8651.
17. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, et al. (2011) Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet 7: e1002197. doi:10.1371/journal.pgen.1002197
18. Fu J, Wolfs MG, Deelen P, Westra HJ, Fehrmann RS, et al. (2012) Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet 8: e1002431. doi:10.1371/journal.pgen.1002431
19. Arakawa S, Takahashi A, Ashikawa K, Hosono N, Aoi T, et al. (2011) Genome-wide association study identifies two susceptibility loci for exudative age-related macular degeneration in the Japanese population. Nat Genet 43: 1001–1004.
20. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996–1006.
21. Kashihara K (2006) Weight loss in Parkinson's disease. J Neurol 253 Suppl 7: VII38–41.
22. Patel M, Chan CC (2008) Immunopathological aspects of age-related macular degeneration. Semin Immunopathol 30: 97–110.
23. Johnstone RW, Frew AJ, Smyth MJ (2008) The TRAIL apoptotic pathway in cancer onset, progression and therapy. Nat Rev Cancer 8: 782–798.
24. George A, Figueredo VM (2011) Alcoholic cardiomyopathy: a review. J Card Fail 17: 844–849.
25. Hunter RJ, Neagoe C, Jarvelainen HA, Martin CR, Lindros KO, et al. (2003) Alcohol affects the skeletal muscle proteins, titin and nebulin in male and female rats. J Nutr 133: 1154–1157.
26. Ward LD, Kellis M (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40: D930–934.
27. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, et al. (2012) Architecture of the human regulatory network derived from ENCODE data. Nature 489: 91–100.
28. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777.
29. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772.
30. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.
31. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529. doi:10.1371/journal.pgen.1000529

4

BIOLOGICAL INTERPRETATION OF GENOME-WIDE ASSOCIATION STUDIES USING PREDICTED GENE FUNCTIONS

Nature Communications, 6:5890, Jan 2015

Tune H. Pers, Juha M. Karjalainen, Yingleong Chan, Harm-Jan Westra, Andrew R. Wood, Jian Yang, Julian C. Lui, Sailaja Vedantam, Stefan Gustafsson, Tonu Esko, Tim Frayling, Elizabeth K. Speliotes, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Michael Boehnke, Soumya Raychaudhuri, Rudolf S. N. Fehrmann, Joel N. Hirschhorn & Lude Franke

Supplementary information is available at https://goo.gl/Ya27v7

nature COMMUNICATIONS

ARTICLE

Received 29 Sep 2014 | Accepted 18 Nov 2014 | Published 19 Jan 2015

DOI: 10.1038/ncomms6890

Biological interpretation of genome-wide association studies using predicted gene functions

Tune H. Pers[1,2], Juha M. Karjalainen[3], Yingleong Chan[1,2,4], Harm-Jan Westra[5], Andrew R. Wood[6], Jian Yang[7,8], Julian C. Lui[9], Sailaja Vedantam[1,2], Stefan Gustafsson[10], Tonu Esko[1,2,11], Tim Frayling[6], Elizabeth K. Speliotes[12], Genetic Investigation of ANthropometric Traits (GIANT) Consortium[†], Michael Boehnke[13], Soumya Raychaudhuri[2,5,14,15,16], Rudolf S.N. Fehrmann[3], Joel N. Hirschhorn[1,2,4,⋆] & Lude Franke[3,⋆]

The main challenge for gaining biological insights from genetic associations is identifying which genes and pathways explain the associations. Here we present DEPICT, an integrative tool that employs predicted gene functions to systematically prioritize the most likely causal genes at associated loci, highlight enriched pathways and identify tissues/cell types where genes from associated loci are highly expressed. DEPICT is not limited to genes with established functions and prioritizes relevant gene sets for many phenotypes.

[1] Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts 02115, USA. [2] Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 2142, USA. [3] Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen 9711, The Netherlands. [4] Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [5] Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [6] Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter EX1 2LU, UK. [7] Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia. [8] The University of Queensland Diamantina Institute, The Translation Research Institute, Brisbane, Queensland 4012, Australia. [9] Section on Growth and Development, Program in Developmental Endocrinology and Genetics, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892, USA. [10] Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala 75185, Sweden. [11] Estonian Genome Center, University of Tartu, Tartu 51010, Estonia. [12] Department of Internal Medicine, Division of Gastroenterology, and Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA. [13] Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA. [14] Partners HealthCare Center for Personalized Genetic Medicine, Boston, Massachusetts 02115, USA. [15] Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [16] Faculty of Medical and Human Sciences, University of Manchester, Manchester M13 9PL, UK. ⋆ These authors contributed equally to this work. Correspondence and requests for materials should be addressed to T.H.P. (email: tunepers@broadinstitute.org) or to J.N.H. (email: joelh@broadinstitute.org) or to L.F. (email: lude@ludesign.nl).

† List of members and affiliations appears as Supplementary Note 4.

The causal variants, genes and pathways in many genome-wide association studies (GWAS) loci often remain elusive, due to linkage disequilibrium (LD) between associated variants, long-range regulation and incomplete biological knowledge of gene function. To translate genetic associations into biological insight, we need at a minimum to identify the genes that account for associations as well as the pathways and tissue/cell type context(s) in which the genes' actions affect phenotypes. Although cell-type-specific expression quantitative trait loci (eQTLs) or coding (non-synonymous) variants in strong LD with associated variants can potentially link these variants to genes, overlap with eQTLs or coding variants may be coincidental. In addition, coding variants in high LD with associated variants are rarely observed, and eQTL data from non-haematological cell types are rare. Direct functional follow-up of the many potentially causal variants and genes is typically difficult and expensive, so an attractive first step is to use computational approaches to prioritize genes in associated loci with respect to their likely biological relevance, and to identify pathways and tissues to define their likely biological context. The current paradigm for gene prioritization methods is to systematically search for commonalities in functional annotations between genes from different associated loci, such as shared features derived from text mining[1] (which is limited by the literature's highly incomplete characterization of gene function) or propensity to interact at the protein level[2] (which is unlikely to capture the full functional spectrum of a given gene or phenotype[3]). The paradigm for gene set analysis is to search for enrichment of the genes near associated variants in manually curated gene sets or in gene sets derived from molecular evidence[4]. Although certain pathways have been carefully characterized, and manually curated gene sets and protein–protein interaction maps can be of great value, pathway annotation of genes remains sparse and skewed towards well-studied genes[5]. At the same time, the availability of large, diverse, genome-wide data sets, such as gene expression data, can elucidate and annotate potential functional connections between genes[6]. Given these limitations and opportunities, and the wide spectrum of traits and diseases analysed in association studies, there is a need for a general computational approach that integrates diverse, non-hypothesis-driven data sets to prioritize genes and pathways[7,8].

With the goal of meeting this need, we develop and hereby present a framework called Data-driven Expression Prioritized Integration for Complex Traits (DEPICT, www.broadinstitute.org/depict), which is not driven by phenotype-specific hypotheses and considers multiple lines of complementary evidence to accomplish gene prioritization, pathway analysis and tissue/cell type enrichment analysis. This framework can prioritize genes, pathways and tissue/cell types across many different phenotypes[9–13].

Results

Overview of the DEPICT methodology. DEPICT builds on our recent work that used co-regulation of gene expression (derived from expression data of 77,840 samples), in conjunction with previously annotated gene sets, to accurately predict gene function based on a 'guilt-by-association' procedure[6]. We first expanded this approach to include 14,461 existing gene sets, representing a wide spectrum of biological annotations (including manually curated pathways[14–16], molecular pathways from protein–protein interaction screens[17] and phenotypic gene sets from mouse gene knock-out studies[18]). By calculating, for each gene, the likelihood of membership in each gene set (based on similarities across the expression data; see Methods), we generated 14,461 'reconstituted' gene sets (see Fig. 1; Supplementary Data 1). Rather than traditional binary gene sets (genes are included or not included), these reconstituted gene sets contain a membership probability for each gene in the genome; conversely, a gene is functionally characterized by its membership probabilities across the 14,461 reconstituted gene sets. Using these precomputed gene functions and a set of trait-associated loci, DEPICT assesses whether any of the 14,461 reconstituted gene sets are significantly enriched for genes in the associated loci, and prioritizes genes that share predicted functions with genes from the other associated loci more often than expected by chance. In addition, DEPICT utilizes a set of 37,427 human microarrays to identify tissue/cell types in which genes from associated loci are highly expressed. DEPICT uses precomputed GWAS based on randomly distributed phenotypes to take sources of confounding into account: it extracts gene-density-matched input loci from these 'null GWAS', recomputes results and adjusts the P values from the above three analyses for null expectation. It also uses the null GWAS to adjust for multiple testing by computing false discovery rates (FDRs, see Methods).

Calibration of locus definitions. Having developed this framework, we first considered a key feature, the definition of an associated locus—that is, given an associated variant, how many of the nearby genes should be taken into consideration as potentially causal? Using as a positive control Mendelian disease genes that affect skeletal growth and are over-represented in height-associated GWAS loci[10,19], we evaluated DEPICT's performance using loci defined by different combinations of genetic and physical distance from the lead associated variant (Supplementary Data 2). We found that a locus definition of $r^2 > 0.5$ from the lead variant was optimal (Supplementary Note 1). We repeated the analysis using genome-wide-significant associations for low-density lipoprotein (LDL) cholesterol[20] and 14 Mendelian lipid genes[20] as positive controls and observed similar results ($r^2 > 0.4$), indicating that the calibration does not change drastically for other traits (Supplementary Data 3).

Type-1 error rate analysis. We next tested whether DEPICT properly controls the type-1 error rate. Running DEPICT with random input loci based on either real genotype or simulated genotype data, we observed nearly uniform distributions for gene set enrichment, gene prioritization and tissue/cell type enrichment P values (see Supplementary Fig. 1 and Methods). Importantly, we did not observe any correlation between gene length and gene prioritization P values (Spearman $r^2 = 7.70 \times 10^{-5}$), nor correlation with locus gene density (Spearman $r^2 = 7.53 \times 10^{-8}$), two factors that have often confounded pathway analyses[21]. We also did not observe any correlation between tissue/cell type enrichment P values and the number of samples available in the expression data sets for each annotation (Spearman $r^2 = 6.9 \times 10^{-4}$), nor were results dependent on the particular set of genotype data used to construct the null GWAS (Supplementary Note 2). Together, these results indicated that DEPICT results are not driven by bias in its data sources.

Benchmarking the gene set enrichment framework. We next compared DEPICT with two GWAS pathway methods, MAGENTA[22] and GRAIL[1] using GWAS results for three phenotypes, each with >50 independent genome-wide significant single-nucleotide polymorphisms (SNPs): Crohn's disease[23], human height[10] and LDL[20]. DEPICT's gene set enrichment functionality outperformed MAGENTA (a widely used GWAS gene set enrichment tool) by identifying more relevant gene sets (both methods exhibited comparable type-1 error rates; Supplementary Figs 1 and 2) for all three phenotypes: DEPICT identified 2.5 times as many significant gene sets (FDR < 0.05) for Crohn's disease, 2.8 times as many significant gene sets for height and 1.1 times as many significant gene sets for LDL (Fig. 2; Supplementary Figs 3–5; Supplementary Data 4–6). Many gene sets prioritized by DEPICT, but not MAGENTA, appear biologically relevant (for example, regulation of immune response, response to cytokine stimulus and toll-like receptor signalling pathway for Crohn's disease; Fig. 2). To test whether our gene set reconstitution strategy was driving the performance differences between MAGENTA and DEPICT, we ran MAGENTA with non-probabilistic, binary (yes/no) versions of the reconstituted gene sets (see Methods). We found a consistent increase in the number of nominally significant gene sets when MAGENTA was run with reconstituted gene sets for Crohn's disease, height and LDL (1.4, 1.6 and 1.7-fold increases, respectively, in number of nominally significant gene sets using the 95 percentile model; Supplementary Data 4–6; Supplementary Figs 6–8). To assess whether the reconstituted gene sets enhance the performance of DEPICT, we ran DEPICT using the original, predefined gene sets. As expected, the number of prioritized gene sets (FDR < 0.05) dropped to 97.7, 92.9 and 20% for the Crohn's disease, height and LDL analyses, respectively (Supplementary Data 4–6). Together, these analysis indicate that the gene set reconstitution, combined with DEPICT's ability to use probabilistic gene sets, is responsible for the increased performance of DEPICT compared with MAGENTA in gene set enrichment analysis.

Benchmarking the gene prioritization framework. Using gene lists from whole-blood expression quantitative locus data[24], rodent growth plate differential expression data[25] and Mendelian human lipid genes reported in literature[20] (see Methods), we constructed positive sets of genes to compare DEPICT's gene prioritization performance with GRAIL (a widely used GWAS gene prioritization tool). DEPICT and GRAIL performed similarly in analyses based on all genome-wide significant loci with at least one positive gene, based on area under a receiver-operating characteristic (ROC) curve (AUC, Table 1; Supplementary Datas 7–9; Supplementary Fig. 9). However, when restricting the height comparison with loci with

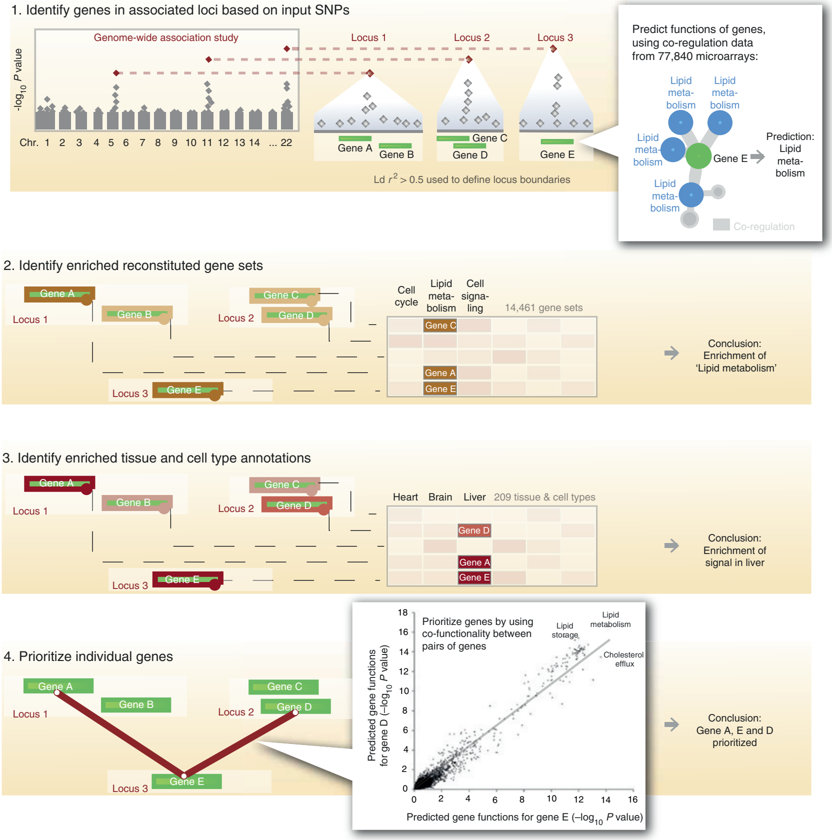

1. Identify genes in associated loci based on input SNPs
Genome-wide association study
−log10 P value
Chr. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 ... 22
Locus 1
Locus 2
Locus 3
Gene A
Gene B
Gene C
Gene D
Gene E
Ld r2 > 0.5 used to define locus boundaries
Predict functions of genes, using co-regulation data from 77,840 microarrays:
Lipid metabolism
Lipid metabolism
Lipid metabolism
Lipid metabolism
Gene E
Prediction: Lipid metabolism
Co-regulation
2. Identify enriched reconstituted gene sets
Gene A
Gene B
Gene C
Gene D
Gene E
Locus 1
Locus 2
Locus 3
Cell cycle
Lipid metabolism
Cell signaling
14,461 gene sets
Gene C
Gene A
Gene E
Conclusion: Enrichment of 'Lipid metabolism'
3. Identify enriched tissue and cell type annotations
Gene A
Gene B
Gene C
Gene D
Gene E
Locus 1
Locus 2
Locus 3
Heart
Brain
Liver
209 tissue & cell types
Gene D
Gene A
Gene E
Conclusion: Enrichment of signal in liver
4. Prioritize individual genes
Gene A
Gene B
Gene C
Gene D
Gene E
Locus 1
Locus 2
Locus 3
Prioritize genes by using co-functionality between pairs of genes
Lipid storage
Lipid metabolism
Cholesterol efflux
Predicted gene functions for gene D (−log10 P value)
Predicted gene functions for gene E (−log10 P value)
0 2 4 6 8 10 12 14 16 18
0 2 4 6 8 10 12 14 16
Conclusion: Gene A, E and D prioritized

Figure 1 | Overview of DEPICT. DEPICT is designed to identify likely causal genes, functional or phenotypic gene sets that are enriched in genes within associated loci, and tissues or cell types that are implicated by the associated loci. DEPICT takes as input a set of trait-associated SNPs and uses them to identify independently associated loci that may comprise up to several genes. DEPICT uses co-regulation data from 77,840 microarrays to predict genes' biological functions across 14,461 gene sets representing a wide spectrum of biological annotations and to construct 14,461 'reconstituted' gene sets. DEPICT then uses this information to identify reconstituted gene sets that enrich for genes in the associated loci, and to prioritize genes at associated loci, by identifying genes in different loci that have similar predicted functions. Finally, DEPICT relies on 37,427 human gene expression microarrays to assess whether genes in associated loci are highly expressed in any of 209 tissue/cell type annotations.

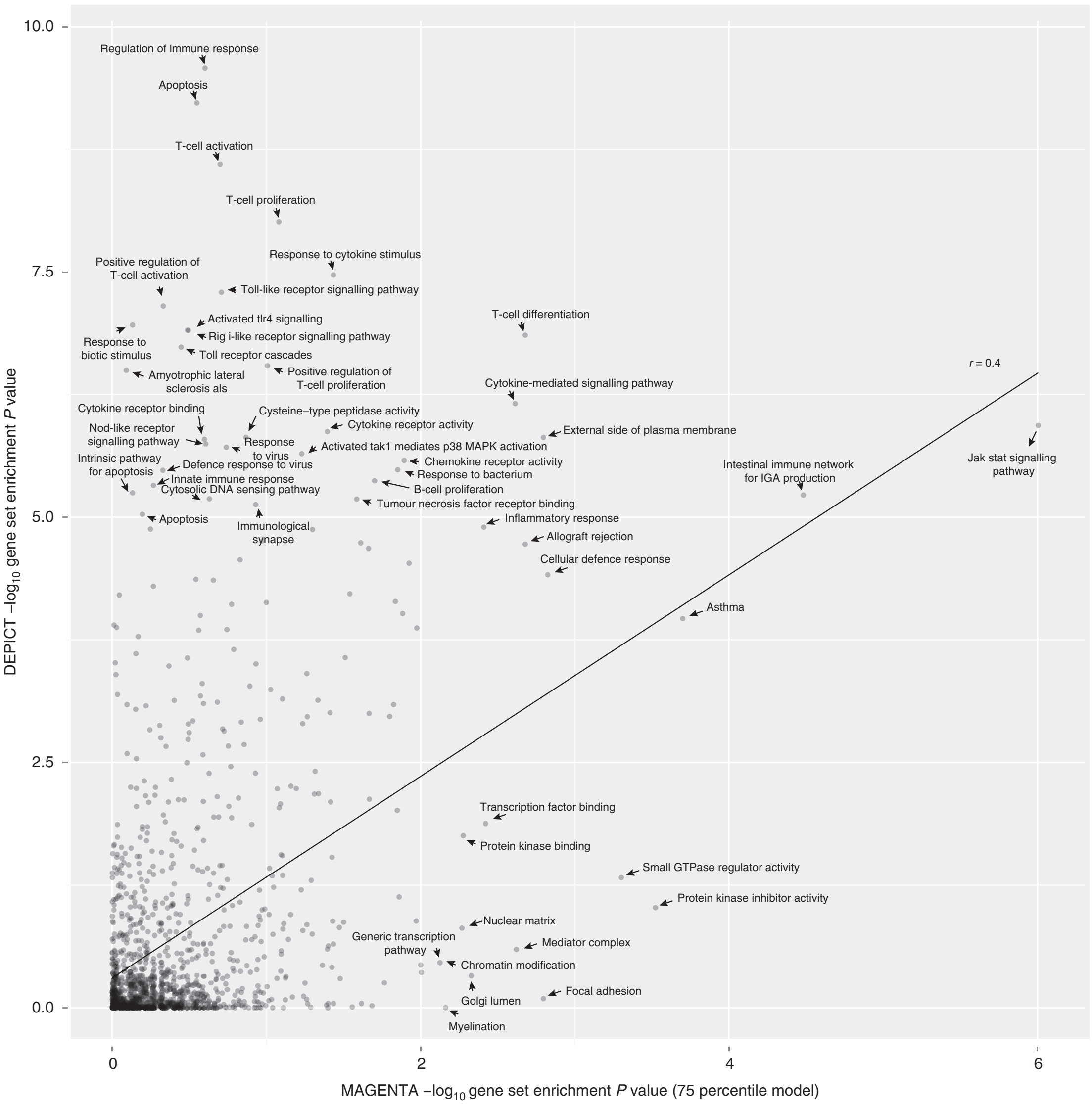


Figure 2 | Comparison of DEPICT and MAGENTA for Crohn's disease. Comparison of DEPICT, which was run with 63 genome-wide significant Crohn's disease SNPs as input, and MAGENTA, which was run using the complete list of Crohn's disease summary statistics[23] (downloaded from www.ibdgenetics.org). DEPICT was run using 1,280 reconstituted gene sets, and MAGENTA was run using the predefined versions of the same 1,280 gene sets. Both methods were run with default settings and non-adjusted enrichment P values are plotted.

no well-known Mendelian human skeletal growth gene, DEPICT markedly outperformed GRAIL, prioritizing genes at many more loci (DEPICT: 1.1 genes per locus, GRAIL: 0.4 genes per locus), suggesting that DEPICT performs better at loci harbouring genes with less well-established roles in literature (Supplementary Data 10). We validated this observation using genes nearest to height-associated SNPs as positive genes at these loci. The nearest gene is an unbiased, but highly imperfect benchmark (for example, only 13/21 Mendelian skeletal growth genes in a large height GWAS[19] were the nearest genes to a height-associated SNP), so AUC is expected to be low using this benchmark. Nonetheless, DEPICT not only prioritized more genes than GRAIL, but also had a higher AUC (Supplementary Data 11). Finally, DEPICT performed consistently better than a gene expression-based version of GRAIL (Supplementary Data 7–9), indicating that use of gene expression data in the prediction is not driving DEPICT's superior performance across several of the comparisons. Together, these analyses indicate that DEPICT performs particularly well for gene prioritization at what are arguably the most important loci for new discovery: those with biology that is less well captured in already published literature.

Prioritization of genes outside genome-wide significant loci. We hypothesized that DEPICT could also be used to prioritize

Table 1 | Overview of DEPICT and GRAIL comparison.

Comparison	Trait/disease	Gold standard genes	Method	ROC AUC	F-measure at $P = 0.05$	Maximum F-measure
Prioriziation at loci with no Mendelian human skeletal growth genes	Human height	Nearest to associated SNPs	DEPICT	0.63	0.66	0.74
			GRAIL	0.60	0.47	0.74
Prioriziation at loci with no Mendelian human skeletal growth genes	Human height	Growth plate biology	DEPICT	0.76	0.82	0.82
			GRAIL	0.57	0.56	0.78
All loci, default method settings	Crohn's disease	eQTLs	DEPICT	0.68	0.60	0.62
			GRAIL	0.71	0.39	0.69
	Human height	Growth plate biology	DEPICT	0.78	0.80	0.82
			GRAIL	0.64	0.59	0.70
	LDL cholesterol	Mendelian lipid disorders	DEPICT	NA	1.00	1.00
			GRAIL	NA	1.00	1.00
All loci, GRAIL with Gene Expression Atlas data	Crohn's disease	eQTLs	DEPICT	0.70	0.62	0.64
			GRAIL (Exp.)	0.68	0.44	0.64
	Human height	Growth plate biology	DEPICT	0.73	0.76	0.78
			GRAIL (Exp.)	0.61	0.44	0.70
	LDL cholesterol	Mendelian lipid disorders	DEPICT	0.83	0.92	0.92
			GRAIL (Exp.)	0.79	0.83	0.92

AUC, area under the curve; DEPICT, Data-driven Expression Prioritized Integration for Complex Traits; eQTLs, expression quantitative trait loci; LDL, low-density lipoprotein; NA, not available; ROC, receiver-operating characteristics curve; SNP, single-nucleotide polymorphism.
DEPICT and GRAIL[1] ROC AUC estimates, and precision and recall estimates for genome-wide significant SNPs for Crohn's disease[23], human height[10] and low-density lipoprotein cholesterol[20]. The height comparison was conducted as loci with and without Mendelian human stature genes[19] to assess which method performed best at loci without known height biology. All comparisons were conducted based on all loci using the default version of GRAIL except the comparisons labelled 'Exp.', which were conducted using GRAIL with Human Gene Expression Atlas data[40] instead of literature. NA because there were the only positive genes in benchmarking loci.

genes outside genome-wide significant loci, based on predicted functional relatedness to genes within genome-wide significant loci. Similar to the gene prioritization implemented in DEPICT, we prioritized genes with higher than expected pairwise similarities to genes from trait-associated loci (across the 14,461 functional predictions; see Methods). SNPs within or near (±50 kb) the 3,022 genes that were functionally related to Crohn's disease loci genes (at $\mathrm{FDR} < 0.05$) had lower association P values than SNPs in the same number of unrelated genes (genes with $\mathrm{FDR} > 0.99$; genomic inflation factor $\boldsymbol{\lambda} = 1.49$ versus $\boldsymbol{\lambda} = 1.31$), indicating that DEPICT enriches for as-yet-unidentified genes associated with Crohn's disease. The enrichment was further increased when considering only SNPs that overlap with eQTLs in whole blood[24] ($\boldsymbol{\lambda} = 1.69$ versus $\boldsymbol{\lambda} = 1.25$). A similar enrichment of associations was seen for height ($\boldsymbol{\lambda} = 1.92$ versus $\boldsymbol{\lambda} = 1.62$) and LDL ($\boldsymbol{\lambda} = 1.06$ versus $\boldsymbol{\lambda} = 0.97$).

To begin to assess the performance and specificity of DEPICT across a wider range of phenotypes, we applied DEPICT to 61 phenotypes in the NHGRI GWAS Catalog[26] that had at least 10 genome-wide-significant (unadjusted association P value $< 5 \times 10^{-8}$) associations. DEPICT identified at least one significantly enriched (P value $< 10^{-6}$, the Bonferroni-corrected significance threshold) reconstituted gene set for 39 of the 61 phenotypes (Fig. 3; Supplementary Data 12). To test whether DEPICT identified similar gene sets for related phenotypes, we clustered the 39 traits based on their gene set enrichment scores across the 14,461 reconstituted gene sets (Fig. 3). Related traits clustered with each other, but different phenotypes yielded quite different gene sets. Furthermore, many of the top gene sets were of clear relevance to the phenotype (Supplementary Data 12). Thus, DEPICT is able to identify, with specificity, biologically relevant gene sets for a wide range of human traits and diseases. Consistent with these results, we recently used DEPICT to analyse GWAS data for height, body mass index and waist–hip ratio adjusted for body mass index (from the GIANT Consortium)[10,12,13] and for hypospadias[9]. For each phenotype, DEPICT highlighted a distinct and biologically meaningful group of known and novel genes, gene sets and tissue/cell types.

Discussion

We present a computational framework called DEPICT, which enables gene prioritization, gene set enrichment analysis and tissue/cell type enrichment analysis to generate specific testable hypotheses that are critical to inform experimental follow-up of GWAS. DEPICT implements these three distinct functionalities into a single, publicly available tool. Apart from providing useful insights into pathways and biological annotations of relevance to a phenotype, a key application of the gene set enrichment functionality is to use it for selecting *in vitro* phenotypes that may serve as readouts in cellular assays used to validate prioritized genes for a complex trait. A key advantage of DEPICT over existing tools is the gene set reconstitution, which enables prioritization of previously poorly annotated genes, as well as more specific and powerful gene set enrichment analysis. By using data sets and methods that are not specific to any particular disease or trait, DEPICT does not depend on phenotype-specific hypotheses (for example, particular neuronal gene sets being important for schizophrenia).

On the basis of our current experience, we recommend employing DEPICT on genome-wide significant loci as well as all loci with association P values $< 10^{-5}$ (see Supplementary Fig. 10 for results based on LDL loci using the relaxed threshold and for an example on visualizing DEPICT results). We also recommend a locus definition of $r^2 > 0.5$ from lead SNPs. It is important to note that reconstituted gene sets should be interpreted in light of the genes that are mapped to them, rather than strictly by their identifiers (which are carried over from the predefined gene sets).

Despite DEPICT's ability to identify relevant gene sets for a large number of traits and diseases, the method may be less sensitive to phenotypes caused by genes that have specialized functions that cannot be well predicted based on integrating gene expression data with the currently existing predefined gene sets. Indeed, there are multiple ways in which the DEPICT framework could be improved further. Additional future work includes iteratively conditioning on significant genes, gene sets and tissue/cell types to enhance prioritization of genes with weaker, yet

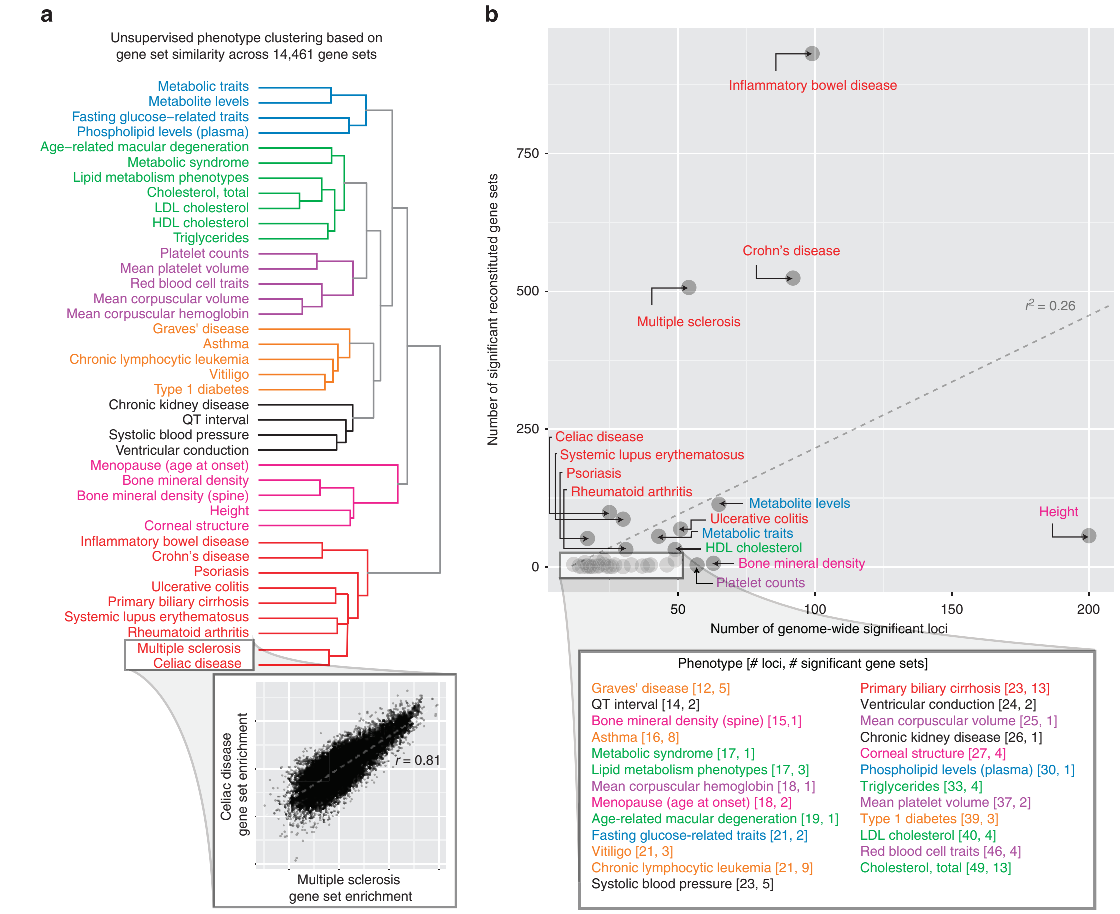


Figure 3 | DEPICT analysis using GWAS Catalog results. DEPICT identified at least one significant reconstituted gene set for 39 traits and diseases from the GWAS Catalog (we investigated 61 traits with at least 10 independent genome-wide significant loci). (**a**) Unsupervised clustering of the 39 phenotypes based on their gene set enrichment scores across all reconstituted gene sets yielded 7 clusters of phenotypes (roughly corresponding to metabolic, lipids, haematological, autoimmune, blood pressure/cardiac conduction, growth/bone/menopause and a second autoimmune cluster), which indicates that DEPICT is able to identify phenotypic-specific and biologically relevant gene sets for a wide range of phenotypes. The inset shows that the multiple sclerosis and coeliac disease gene set enrichment scores are highly correlated and therefore were clustered within the same clade. (**b**) The number of genome-wide significant loci for a given phenotype was positive correlated with the number of significant (FDR < 0.05) reconstituted gene sets for that phenotype (Pearson $r^2 = 0.26$, *t*-test *P* value $= 6.86 \times 10^{-5}$).

significant, relationships, and quantification of the relative importance of significant predictions. Additional expression data would enhance the data sources available for DEPICT, especially for prioritization of tissues/cell types. Other data types, such as epigenetic data, have yet to be integrated into the DEPICT framework, and DEPICT does not yet use information that could further prioritize genes within loci, such as LD with eQTLs or missense variation, or being the nearest gene to the lead SNP. Finally, DEPICT is currently optimized for GWAS results, but could be adapted to other types of data sets (custom arrays, exome chip or sequencing).

In conclusion, there is a need for approaches that are not driven by phenotype-specific hypotheses and that consider multiple lines of complementary evidence to accomplish gene prioritization, pathway analysis and tissue/cell type enrichment analysis. We have developed a computational and publicly available tool—DEPICT—that can address this need by performing integrative analysis, thereby generating novel, testable hypotheses from genetic association studies across a wide spectrum of traits and diseases.

Methods

Data and software availability. The following sections describe the DEPICT methodology in detail. DEPICT source code and example data are available at https://github.com/DEPICTdevelopers. Ready-to-use software is available at www.broadinstitute.org/depict.

Definition of associated loci. From the set of associated SNPs at a particular threshold (such as genome-wide significance, $P < 5 \times 10^{-8}$), we generated independent 'lead SNPs' by retaining the most significant SNP from each set of SNPs that are in LD (pairwise $r^2 > 0.1$) and/or in proximity (physical distance of < 1 Mb). We computed pairwise LD coefficients based on the imputation panel used in the GWAS, either HapMap Project release 2 and 3 CEU genotype data[27] or 1000 Genomes Project Phase 1 CEU, GBR and TSI genotype data[28]. We defined positions in the human genome according to genome build GRCh37. Next, we created lists of genes at associated loci by mapping genes to loci if they resided within, or were overlapping with, boundaries defined by the most distal SNPs in either direction with LD $r^2 > 0.5$ to the given lead SNP (see Supplementary Note 1 for justification of this locus definition). If no genes were within the locus defined by $r^2 > 0.5$, the gene nearest to the given lead SNP was included. Loci with overlapping genes were then merged. Due to the extended LD in the major histocompatibility complex region and the resulting challenges in delineating associated loci, genes within base pairs 25,000,000–35,000,000 on chromosome 6 were excluded. DEPICT takes as input a set of independent, associated SNPs and automates all other steps outlined here.

Gene sets used in DEPICT. DEPICT is based on a large number of predefined gene sets from diverse databases and data types (Supplementary Data 1). Gene ontology[15], Kyoto encyclopedia of genes and genomes[14] and REACTOME[16] gene sets were mapped to Ensembl database identifiers. Molecular pathways were constructed based on experimentally derived high-confidence protein–protein interactions from the InWeb database[17] by considering each of the 12,793 genes in the database and annotating direct, high-confidence interaction partners of a given gene as a molecular pathway (including the given gene itself). We defined high-confidence interactions as pairs of gene products with InWeb-specific protein–protein interaction confidence scores above 0.154, a cutoff previously justified[17]. In addition, we constructed 2,473 phenotypic gene sets based on 211,882 phenotype–gene relationships from the Mouse Genetics Initiative[18]. These gene sets were constructed by adding genes to the same gene set if they were related to the same Mouse Genetics Initiative phenotype. From all repositories, we only included gene sets with at least 10 genes and at most 500 genes.

Gene function prediction for gene set reconstitution. DEPICT performs gene prioritization and gene set enrichment based on predicted gene function and reconstituted gene sets (note that the reconstituted gene sets are a consequence of the gene function prediction). Please refer to Fehrmann *et al.*[6] (and www.genenetwork.nl) for a detailed description of the gene function prediction method. The main hypothesis behind the gene function prediction follows a guilt-by-association logic: a gene that is co-regulated with say 20 other genes, which perform a specific function, is likely to exhibit the same function. In Fehrmann *et al.*[6], we developed an approach that quantifies co-regulation between pairs of genes based on gene expression data, even in instances where transcriptomic co-regulation is subtle. In Fehrmann *et al.*[6], we conducted the following steps to predict functions of genes and construct reconstituted gene sets:

1. We first renormalized 77,840 microarrays from two human, one rat and one mouse Affymetrix gene expression platform downloaded from the Gene Expression Omnibus (GeO) database[29] (Supplementary Data 13).
2. We constructed a probe–probe correlation matrix (using Pearson correlation to compute all pairwise probesets correlations) for each of the four platforms.
3. We performed principal component analysis on each of the four correlation matrices, and used Cronbach's Alpha and Split-half reliability statistics to retain 777 and 377 eigenvectors (hereafter 'transcriptional components' or 'TCs'; Fehrmann *et al.*[6]) from the two human platforms, 677 TCs from the mouse platform and 375 TCs from the rat platform.
4. We mapped all human genes to Ensembl identifiers[30]; mouse and rat genes were mapped to their human homologues (Ensembl database orthology mapping). The loadings of each gene on each TC are the elements of a gene-TC matrix with 19,997 gene rows (the number of genes covered by the Affymetrix platforms) and 2,206 TC columns.

We then used the gene-TC matrix to predict 19,997 genes' function across the 14,461 functional annotations represented by the predefined gene sets, by doing the following steps:

1. For each gene set, we computed the enrichment on each TC (using z-scores derived from Welch's t-test to assess whether the TC loadings from genes from the given set significantly deviated from all other genes' loadings). This resulted in a TC profile for each gene set (a gene set-TC matrix of z-scores with 14,461 gene set rows and 2,206 TC columns).
2. To obtain gene function predictions and reconstituted gene sets, we quantified each gene's likelihood of being part of a given predefined gene set by correlating the gene's 2,206 TC loadings (from the gene-TC matrix) with the z-score TC profile of each gene set (from the gene set-TC matrix). To avoid circularity in cases where a particular gene was part of a predefined gene set, we left out that gene from the gene set, recomputed the gene set z-score profiles along all TCs and then computed the correlation of the gene with the gene set.
3. We converted the correlation P values to z-scores to obtain a gene-gene set matrix of z-scores comprising 19,997 gene rows and 14,461 gene sets columns. This matrix is used by DEPICT to perform gene prioritization and gene set enrichment analysis.

Null GWAS construction. To take sources of confounding into account, DEPICT makes use of precomputed GWAS based on randomly distributed phenotypes to ('null GWAS'). We computed 200 GWAS based on genome-wide CEU genotype data from the Diabetes Genetics Initiative[31] (DGI) and simulated Gaussian phenotypes (random draws from $N(0,1)$ distribution) with no genetic basis.

DEPICT gene prioritization. For gene prioritization, DEPICT employs a phenotype- and mechanism-agnostic algorithm, which is predicated on a previously formulated assumption that truly associated genes share functional annotations[1,17,32]. In other words, genes within associated loci that are functionally similar to genes from other associated loci are the most likely causal candidates. DEPICT prioritizes genes based on three major steps: a scoring step, a bias adjustment step and a FDR estimation step. In the scoring step, the method quantifies the similarity of a given gene to genes from other associated loci by correlating their reconstituted gene set memberships (across all 14,461 gene sets). The bias adjustment step is designed to control inflation in gene scores caused by, for example, gene length (longer genes are more likely to be part of associated GWAS loci) or structure in the underlying expression data. In this step, the method normalizes the given gene's similarity score based on the distribution of the given gene's similarity to genes from 1,000 sets of gene-density-matched loci, derived from the 200 pre-permuted null GWAS. In the last step, experiment-wide FDRs are estimated by repeating the scoring and bias adjustment steps 20 times based on top SNPs from precomputed null GWAS. For a given gene (gene x) that has a prioritization P value y in the actual data, a FDR is calculated by first counting the number of genes having a P value smaller or equal to y across all 20 null runs and dividing this count by the rank of gene x in the actual data. We note that in the version of DEPICT implemented in the studies of anthropometric traits[10,12,13], we included a correction for the number of genes at a given locus. Because this correction does not change gene prioritization results markedly (gene set enrichment results and tissue/cell type enrichment results are unchanged), we recommend not using this correction because it imposes an overly conservative correction on genes in relatively gene-poor loci. This correction was not implemented in the version described here.

DEPICT reconstituted gene set enrichment. The gene set enrichment analysis algorithm comprises the same three steps as employed in gene prioritization: a gene set scoring step, a bias correction step and a FDR estimation step. For a given reconstituted gene set, DEPICT quantifies enrichment by (1) summing the given gene set membership z-scores (entries in the gene-gene set matrix) of all genes within each associated locus and then computing the sum of sums across all loci; (2) repeating step 1 a thousand times based on random loci that are matched by gene density, and using the thousand null z-scores to adjust the real z-score by subtracting their mean, dividing by their s.d. and converting the adjusted z-score to a P value; and (3) repeating steps 1 and 2 twenty times to estimate experiment-wide FDRs similar to the method described above.

DEPICT tissue/cell type enrichment analysis. DEPICT utilizes 37,427 human Affymetrix HGU133a2.0 platform microarrays (approximately half of the microarrays used to reconstituted gene sets) to assess whether genes in associated loci are highly expressed in any of the 209 Medical Subject Heading (MeSH) tissue and cell type annotations. The tissue/cell type expression matrix was constructed by averaging gene expression levels of microarray samples with the same MeSH annotation[6]. This process included $N(0,1)$ normalizing across all tissue/cell type annotations to remove effects of ubiquitously expressed genes, $N(0,1)$ normalizing the columns of the tissue/cell type expression matrix (to allow enrichment analysis identical to the gene set enrichment analysis framework) and retaining only tissue/cell type annotations covered by at least 10 microarrays. Conceptually, the resulting gene-tissue/cell type expression matrix resembles the gene-gene set matrix, the only difference being that columns represent the relative expression of genes in a given tissue compared with the other tissues, as opposed to the likelihood of membership of a gene in a gene set. Consequently, the tissue/cell type enrichment analysis algorithm is conceptually identical to the gene set enrichment analysis algorithm.

Adjusting for confounding sources. For a given set of associated loci from the 'real GWAS' (the study of interest), DEPICT extracts the same number of independent loci from the 200 precomputed null GWAS. For a given null GWAS, this is accomplished by varying the SNP association P value cutoff until the number of independent top loci is the same as the number of independent loci in the real GWAS. The independent top loci from each null GWAS are then collected into a single pool of loci. During the DEPICT gene prioritization, gene set enrichment and tissue/cell type enrichment analyses, this pool of loci is used to sample 1,000 collections of gene density-matched 'null loci' (in each collection there are as many null loci as the number of loci observed in the real GWAS). Null loci within a given collection are not allowed to overlap (in terms of genes). During the DEPICT background correction step, if a locus from the real GWAS is represented by <10 gene-density-matched null loci, DEPICT iteratively includes larger and smaller null loci (to avoid oversampling the same null loci during the 1,000 background runs). We employed different numbers of null GWAS contributing to the pool of null loci, and observed no major differences between using 200, 500 or 900 null GWAS (Supplementary Note 3).

Type-1 error rate analyses. To compute type-1 error rates for the gene prioritization, gene set enrichment and tissue/cell type enrichment analyses, we first computed 100 DGI null GWAS the same way as describe in the above section. Spearman correlation coefficients were computed based on $\log_{10}$ transformed

P values. We used an alternate approach to estimate type-1 error by replacing the null GWAS with simulated GWAS that have positive signals but no underlying biological basis. We simulated 50,000 individuals using HAPGEN[33] using parameters from the HapMap Project release 3 CEU population. From this, we obtained 1,175,577 genotypes for all autosomes (chromosomes 1–22) and calculated the allele frequency for each SNP using the 50,000 individuals. We then randomly selected 1,000 SNPs to have an effect on the phenotype and assigned effect sizes such that all SNPs jointly explain 45% of the total variance. The effect size for each SNP was calculated as follows,

$$\beta = \delta\sqrt{\frac{\sigma^2}{2p(1-p)}} \qquad (1)$$

where β is the effect size in s.d. units, σ^2 is the variance explained for each SNP, p is the SNP's minor allele frequency and δ denotes a random variable with equal probability of being $+1$ or -1. Once each SNP's effect size was determined, we calculated the weighted allele score for each individual by summing up the SNP minor allele dosages weighted by their effect size. The weighted allele score was calculated as follows,

$$\mathrm{WAS} = \sum_{i=1}^{N} \beta_i \mathrm{SNP}_i - 2\beta_i p_i \qquad (2)$$

where N is the number of SNPs ($N = 1{,}000$), β_i is the effect size of the ith SNP as calculated earlier, SNP_i is the dosage of the minor allele for the ith SNP (0,1 or 2) and p_i is the minor allele frequency of the ith SNP. The subtraction of $2\beta_i p_i$ served to adjust the weighted allele score such that its mean was 0. We obtained the final phenotypic *z*-score by adding a remaining noise term such that the total variance was 1. The *z*-score was calculated as follows,

$$z\text{-score} = \mathrm{WAS} + N(0,\ \text{variance\_remaining}) \qquad (3)$$

where *N*(0, variance_remaining) is a randomly generated number sampled from a Normal (*N*) distribution with mean 0 and variance 0.55. This process was repeated 100 times to obtain 100 sets of phenotypic *z*-scores for each of the 50,000 individuals. We used PLINK[34] to perform GWAS on each set of phenotypes using the 50,000 simulated genotype samples, and then, for each null GWAS, identified the association *P*-value threshold that resulted in 100 fully independent loci (DEPICT locus definition). Finally, we ran DEPICT with default settings on each of the $n = 100$ sets of input SNPs.

Crohn's disease DEPICT analysis. Summary statistics from GWAS-based meta analysis of Crohn's disease[23] (downloaded from www.ibdgenetics.org) were used to identify genome-wide significant loci (using PLINK and parameters '--clump-kb 1000 --clump-r2 0.01'). As input to DEPICT we used the resulting 63 genome-wide significant (χ^2-test *P* value $< 5 \times 10^{-8}$), which were located in 54 fully independent loci based on DEPICT definitions of independence.

Human height DEPICT analysis. As input we used 697 genome-wide significant human height associations identified in GWAS-based meta analysis[10] (accessible through http://www.broadinstitute.org/collaboration/giant), which were located in 566 fully independent loci based on DEPICT definitions of independence.

Low-density lipoprotein cholesterol DEPICT analysis. Summary statistics from GWAS-based meta analysis of LDL[20] (downloaded from www.sph.umich.edu/csg/abecasis/public/lipids2010) were used to identify genome-wide significant loci (using PLINK with parameters '--clump-kb 1000 --clump-r2 0.01'). As input to DEPICT we used the resulting 67 independent loci, which resulted in 40 fully independent loci used DEPICT definitions of independence.

Gene set enrichment benchmark. Due to the lack of an unbiased set of gold standard pathways for any complex trait, we compared DEPICT and MAGENTA[22] by counting the number of statistically significant gene sets predicted based on Crohn's disease, height and LDL loci. Prior to the benchmark, we estimated the type-1 error rate of both methods by running them with summary statistics from 100 null GWAS constructed based on simulated Gaussian phenotypes with no genetic basis, and HapMap Project release 2 imputed DGI Consortium genotype data (Supplementary Figs 1 and 3). For the null analyses, the top 200 independent loci from each null GWAS were used as input, whereas genome-wide significant loci were used as input in the Crohn's disease, height and LDL analyses. All MAGENTA runs were based on the complete set of summary statistics. We restricted the comparison to a list of 1,280 gene sets (gene ontology terms, Kyoto encyclopedia of genes and genomes and REACTOME pathways) with overlapping identifiers between both methods. DEPICT was run on reconstituted gene sets. MAGENTA was run with default settings and both methods excluded the major histocompatibility complex region. The non-probabilistic, binary (yes/no) version of the reconstituted gene sets used in one of the MAGENTA comparisons were constructed by applying a threshold on the gene scores for a given reconstituted gene set (all genes above a permutation-based cutoff were considered part of the given reconstituted gene sets, as reported in ref. 6). Entries with 'NA' in columns 'DEPICT with predefined gene sets *P*' and 'DEPICT with predefined gene sets FDR' in Supplementary Data 4–6 marked predefined gene sets for which enrichment could not be computed in the DEPICT analysis based on predefined gene sets.

Gene prioritization benchmark. We ran each method (DEPICT and GRAIL[1]) using their default settings on all genome-wide significant Crohn's disease[23], height[10] and LDL[20] associations. To evaluate the methods' performance on the same set of positive genes (genes that are highly likely to be causal to the phenotype) and negative genes (genes that are unlikely to be causal), we limited the comparison to loci at which there was at least one positive gene present across both methods, and discarded any genes at these benchmark loci that were not considered by each method. For the Crohn's disease comparison, we used as positives 31 genes that were transcriptionally regulated in whole blood[24] by a genome-wide significant Crohn's disease association or a SNP in high LD ($r^2 > 0.7$) with a genome-wide significant SNP. For the height comparison, we used as positives a set of 44 genes that were within genome-wide significant height-associated loci and differentially expressed in rodent growth plate expression studies; we have previously shown that the rodent gene expression data are enriched for genes in height-associated loci[25] (Supplementary Table 2 in Lango Allen *et al.*[19]). For the LDL comparison, we used as positives a set of seven genes with reported Mendelian mutations proposed to cause lipid-related traits[20]. For all three benchmarks, we removed negative genes that had a missense variant in strong LD ($r^2 > 0.7$) with an associated SNP; for the height and LDL benchmarks, we removed negative genes that were transcriptionally regulated[24] by a SNP in strong LD ($r^2 > 0.7$) with an associated SNP; in the height benchmark, we removed negative genes that were differentially expressed in rodent growth plates versus other tissues, spatially regulated across different growth plate zones (hypertrophic versus proliferating, and proliferative versus resting) or temporally regulated in growth plates between week 12 and week 3 at nominal significance in reference[25], and genes that were reported in the high-confidence list in ref. 19. After these steps, we were able to use 42 negative genes across 18 loci as Crohn's disease benchmarks and 37 negative genes across 43 loci as height benchmarks. There were no negative genes among the seven LDL benchmark loci. Positive and negatives genes, are listed in Supplementary Data 7–9. Precision (the fraction of positive genes among all prioritized genes at a given *P*-value threshold) and recall (the fraction of correctly classified positive genes at a given *P*-value threshold also referred to as sensitivity) estimates were used to measures accuracy and summarized using the F-measure, which incorporates the ability to recall positive genes with a high precision into a single measure. (Maximum precision implies no false positives, whereas maximum recall implies no false negatives.) To measure the ability to discriminate positive and negative genes at a relative scale, we also computed ROC AUC estimates. To avoid circularity, the growth plate data[25] and the eQTL data[24] were not part of the data used by any of the three methods tested. The R software[35] and the ROCR R library[36] were used to construct the precision recall and ROC curves and the AUC estimates.

Prioritizing genes outside genome-wide significant loci. To enable prioritization of genes below the genome-wide significance threshold, we scored each gene outside the genome-wide significant loci with respect to its similarity to genes within associated loci. For a given gene outside genome-wide significant loci, we (1) correlated (Pearson) its predicted functions across all 14,461 gene sets to every gene in each of the trait-associated loci, (2) kept the lowest correlation *P* value from each genome-wide significant locus, (3) converted the *P* values to *z*-scores and (4) summed the *z*-scores and converted the sum back to a *P* value (alternative hypothesis: gene functionally related to genes in trait-associated loci). We computed FDRs, by redoing steps 1–4 based loci from null GWAS. Using FDR < 0.05 as the threshold, we identified 3,022, 5,916 and 1,901 related genes for Crohn's disease, height and LDL. For each of the three traits, we then calculated genomic inflation factors for SNP *P* values in the functionally related genes and for SNP *P* values in the same number of genes exhibiting the highest (non-significant) FDRs. We added 50 kb flanking loci to gene boundaries (defined by the boundaries of the most extreme transcripts) and required genes to be at least 1 Mb away from the nearest genome-wide significant locus.

GWAS catalog analysis. The GWAS Catalog[26] was downloaded from www.genome.gov/gwastudies/ (download date: 02 January 2014) and 61 phenotypes with at least 10 fully independent regions (DEPICT definitions) based on genome-wide associations were retained. Hierarchical clustering implemented in the R software method 'hclust' was run with default settings (method = 'complete-linkage', dist = 'euclidean'). The DEPICT locus definitions for all GWAS catalog traits can be downloaded from www.broadinstitute.org/mpg/depict.

Overlap of gene sets and visualization. A previous version of DEPICT used in analyses of anthropometric traits[10,12,13] computed gene set overlap by imposing a threshold on which genes belong to a given reconstituted gene set and then used the Jaccard index to compute pairwise overlaps. Overlapping reconstituted gene sets were grouped as pathway families. Here, we instead computed the pairwise Pearson correlation between all reconstituted gene sets and then used the Affinity Propagation method[37] to group similar reconstituted gene sets. We named each cluster ('meta gene set') by the name of the representative gene set automatically identified by the Affinity Propagation method (for examples, see the top 10 gene set enrichment meta gene sets for Crohn's disease, height and LDL in Supplementary Data 14–16). The R software[35] and a R version of the Affinity Propagation method[38] was used setting the parameters 'maxits' to 10,000 and 'convits' to 1,000 to ensure conversion when thousands of reconstituted gene sets needed to be clustered. We visualized the overlap between pathway families pathways using Cytoscape[39].

References

1. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5,** e1000534 (2009).
2. Leiserson, M. D. M., Eldridge, J. V., Ramachandran, S. & Raphael, B. J. Network analysis of GWAS data. *Curr. Opin. Genet. Dev.* **23,** 602–610 (2013).
3. Moreau, Y. & Tranchevent, L.-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.* **13,** 523–536 (2012).
4. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11,** 843–854 (2010).
5. Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E. & Blake, J. A. On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report. *PLoS Comput. Biol.* **8,** e1002386 (2012).
6. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **46,** 1173–1186 (2014).
7. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21,** 1109–1121 (2011).
8. Pers, T. H. *et al.* Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. *Genet. Epidemiol.* **35,** 318–332 (2011).
9. Geller, F. *et al.* Genome-wide association analyses identify variants in developmental genes associated with hypospadias. *Nat. Genet.* **46,** 957–963 (2014).
10. Wood, A. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).
11. Van der Valk, R. J. P. *et al.* A novel common variant in DCST2 is associated with length in early life and height in adulthood. *Hum. Mol. Genet.* 1–14 (2014).
12. Shungin, D. *et al.* New genetic loci link adipocyte and insulin biology to body fat distribution. *Submitted.*
13. Locke, A. *et al.* Large-scale genetic studies of body mass index provide insight into the biological basis of obesity. *Submitted.*
14. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40,** D109–D114 (2012).
15. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
16. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39,** D691–D697 (2011).
17. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25,** 309–316 (2007).
18. Blake, J. A. *et al.* The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.* **42,** D810–D817 (2014).
19. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (2010).
20. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
21. Raychaudhuri, S. *et al.* Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet.* **6,** e1001097 (2010).
22. Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6,** e1001058 (2010).
23. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).
24. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
25. Lui, J. C. *et al.* Synthesizing genome-wide association studies and expression microarray reveals novel genes that act in the human growth plate to modulate height. *Hum. Mol. Genet.* **21,** 5193–5201 (2012).
26. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
27. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
28. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
29. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41,** D991–D995 (2013).
30. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–D755 (2014).
31. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316,** 1331–1336 (2007).
32. Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* **78,** 1011–1025 (2006).
33. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27,** 2304–2305 (2011).
34. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
35. Ihaka, R. & Gentleman, R. R. A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5,** 299–314 (1996).
36. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21,** 3940–3941 (2005).
37. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315,** 972–976 (2007).
38. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27,** 2463–2464 (2011).
39. Saito, R. *et al.* A travel guide to Cytoscape plugins. *Nat. Methods* **9,** 1069–1076 (2012).
40. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101,** 6062–6067 (2004).

Acknowledgements

T.H.P. was supported by The Danish Council for Independent Research Medical Sciences (FSS) The Alfred Benzon Foundation. J.C.L. was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH. L.F. was financially supported by grants from the Netherlands Organization for Scientific Research (NWO-VENI grant 916-10135 and NWO VIDI grant 917-14374) and a Horizon Breakthrough grant from the Netherlands Genomics Initiative (grant 92519031). The research leading to these results has received funding from the European Community's Health Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 259867. We thank the DGI Consortium for making raw genotype and phenotype data available, and the Global Lipids Genetics Consortium and the International Inflammatory Bowel Disease Genetics Consortium for making summary statistics available. We thank Drs. Ayellet V. Segrè, Elizabeth J. Rossin, Jeffrey Baron, Kasper Lage and Pascal Timshel for helpful comments and discussions. This work was supported by The National Institute of Diabetes and Digestive and Kidney Diseases [2R01DK075787 to J.N.H.].

Author contributions

Planning and design was performed by T.H.P., J.M.K., J.N.H. and L.F. Computational analyses were performed by T.H.P., J.M.K., Y.C., H.-J.W. and L.F. The manuscript was written by T.H.P., J.N.H. and L.F. with relevant comments and suggestions by all co-authors.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* 6:5890 doi: 10.1038/ncomms6890 (2015).

5

DISCUSSION AND FUTURE PERSPECTIVES

Liikaa
taas
tunnen
karmivan
kriitikon
ahdistavan

Ismo Alanko — Taiteilijaelämää

5.1 DISCUSSION

This thesis had two main aims: (1) to present a gene expression-based method to predict novel functions for human genes, and (2) to present a follow-up method to suggest which genes and pathways may be relevant to different phenotypes, based on findings from genome-wide association studies.

5.1.1 GENE FUNCTION PREDICTION

In chapter 2, we reanalysed tens of thousands of high-quality human, mouse and rat microarrays and leveraged principal component analysis (PCA) to determine hundreds of biologically meaningful "transcriptional components". We corrected expression levels of cancer samples by removing variation based on transcriptional components of non-cancer samples. We found that it was possible to identify deletions and duplications in cancer samples based on the residual expression.

To achieve the first aim of this thesis, we also used the transcriptional components to predict gene function. Treating each of these components equally, we diminished the effects of physiological or metabolic factors on gene expression and increased the effects of subtler, but still biologically relevant, factors. Then, we used established knowledge of biological pathways as gene sets to determine the extent to which each transcriptional component affects each pathway. In this way, we created a transcriptional "pathway profile" for each pathway and could then correlate each gene against each pathway profile. See figure 5.1 for an illustration of this method. The end result of this prediction procedure was a score for every gene-pathway combination. We considered that these scores would indicate how likely it is that any given gene is relevant for any given pathway, based on the gene expression data and pathway gene sets used. We used a permutation-based false discovery rate approach to assess the significance of each prediction score.

This gene function prediction method and the accompanying website (genenetwork.nl/genenetwork) have been widely used to predict functions for genes of interest. According to Google Analytics data, the website still has more than a hundred active users every month.

As an example of the use of our gene function predictions, in 2012, the gene responsible for the Vel blood group, *SMIM1*, was found by analysing the exomes of individuals who were negative for the Vel antigen.[1] Most Vel-negative individuals were homozygous for a 17-nucleotide frameshift deletion in *SMIM1*. In the same study, a SNP in *SMIM1* was found to influence the mean haemoglobin concentration of red blood cells. Our top-predicted Gene Ontology biological process for *SMIM1* was "haemoglobin metabolic process" ($p = 1.3 \times 10^{-16}$). Our predictions were used as computational evidence in the study. As an example of a social-science phenotype, we also predicted relevant pathways for candidate genes found in a GWAS of educational attainment.[2] For the *BSN* and *LRRN2* genes, we predicted "neurotransmitter secretion" ($p = 1.0 \times 10^{-30}$) and "synapse organisation" ($p = 2.5 \times 10^{-9}$) as the top Gene Ontology biological processes, respectively.

Our gene function prediction method is solely based on gene expression data whereas other methods, such as GeneMANIA, combine networks from different organisms and data sources, including shared protein domains and physical interactions.[3,4] The STRING database integrates information of both physical and functional protein-protein interactions.[5] However, the relevant predictions and widespread use of our expression-based method are due to several advantages that it has compared to previously suggested guilt-by-association approaches.

First, the gene expression data we used was, to our knowledge, the largest expression data set — 77,840 human, mouse and rat microarrays — ever used for the purpose of predicting gene function. This large amount of data allowed us to extract more biologically meaningful information than had previously been possible. In total, we found 2206 transcriptional components, each of which was enriched for at least one functional gene set.

Second, by treating each transcriptional component equally, we could increase the relative importance of subtle biological effects on gene expression. Since the first components explain most of the expression variance, their effects (such as physiological or metabolic phenomena, or cell cycle) overshadow the effects of later components. However, since the later components — that only explain a small proportion of variance — were also biologically enriched, scaling the components equally allowed us to incorporate more subtle and non-prevalent biological processes into the gene function prediction approach.

Third, we used a "leave-one-out" cross-validation method when calculating the score for a gene towards a pathway to which it is already known to belong. In many approaches the data are split into a training set and validation set to assess the accuracy of the method. However, our leave-one-out method enabled us to systematically calculate scores for all gene-pathway combinations in a fair way: for every gene known to belong to a pathway, we pretended it did not belong

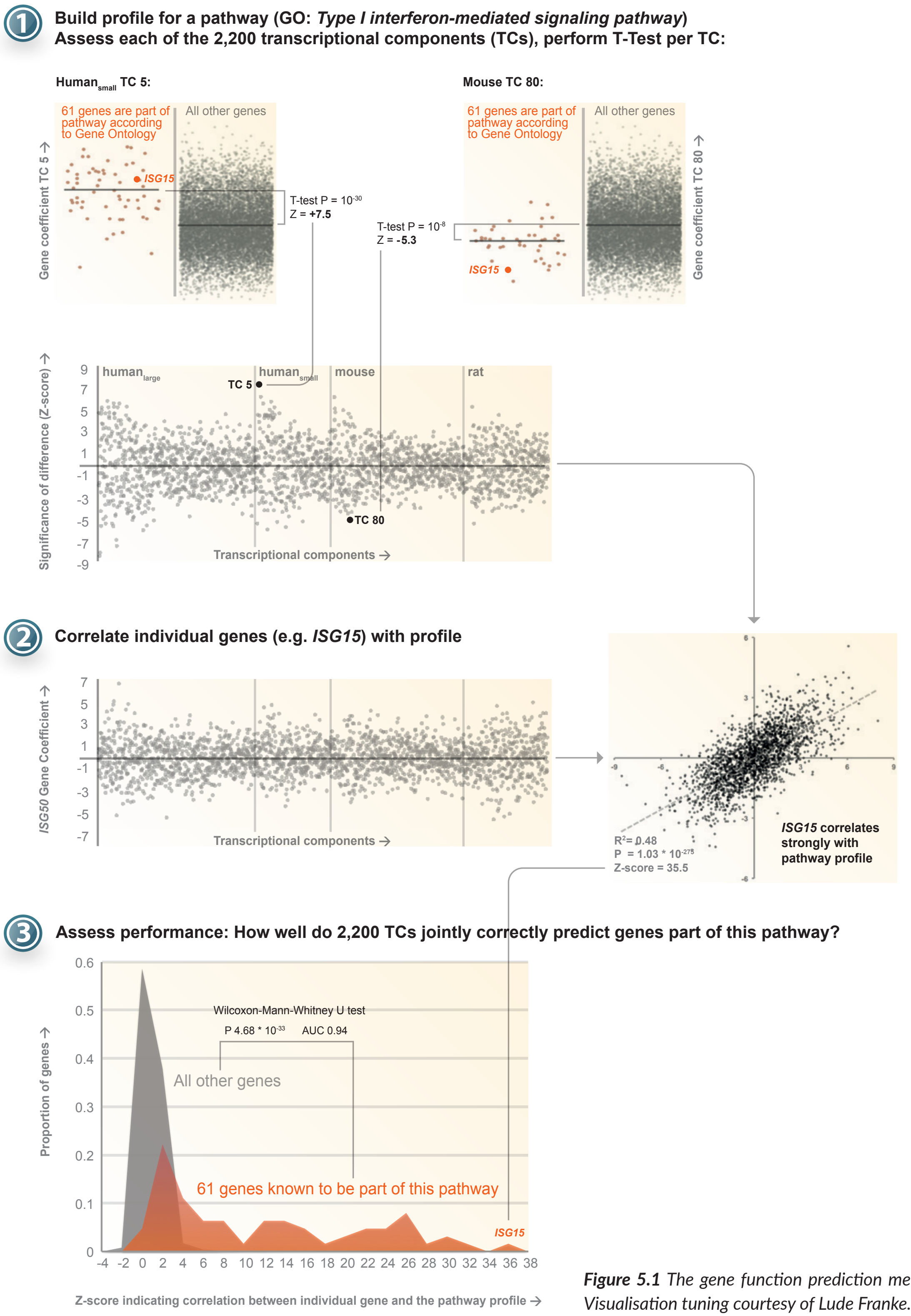


Figure 5.1 *The gene function prediction method. Visualisation tuning courtesy of Lude Franke.*

to that pathway when predicting the gene's functions. Although this increased the computation time of our method dramatically, it could still calculate prediction scores for tens of thousands of genes and pathways, or for other gene sets, using hundreds of transcriptional components, in a matter of hours on a standard laptop.

Fourth, our prediction method not only gives a binary classification of genes to pathways, it also calculates a score for each gene-pathway combination that indicates the strength of that particular prediction. We further leveraged these scores in the DEPICT method described in chapter 4. DEPICT considers genes to be similar if their predicted functions are correlated and uses this similarity measure to prioritise those genes from GWAS loci that share functionality.

Although our gene prediction method has proven useful, there are still several possibilities for improvement.

First, the approach of treating a pathway as merely a gene set, and utilising co-expression or co-regulation patterns across a heteregeneous variety of samples to predict which other genes are similar to that set, works very well for some, but not all, pathways. For example, for the "Unwinding of DNA" and "DNA strand elongation" Reactome pathways, our method gives area under the curve (AUC) values of 0.999 and 0.997, respectively. The high prediction accuracy may be due to the prevalence of DNA replication: it happens in all cells, so strong co-expression patterns of the genes responsible for the process can easily be found when analysing large numbers of samples.

On the other hand, for the "Transport of organic anions" Reactome pathway, our method gives an AUC value of 0.424. This is the only pathway for which the AUC is below 0.5. On average, a random prediction would yield an AUC of 0.5. The poor prediction performance for this pathway may be due to different regulation in different tissues within the *SLCO* gene family that is involved in the pathway.[6] This may hinder our ability to observe global co-expression or co-regulation patterns of these genes. Therefore, for certain pathways, it may be beneficial to restrict the analysis to relevant tissues, or to develop a method that combines the different regulation patterns from different tissues instead of investigating all tissues at once.

Second, despite the usefulness of gene expression data, other layers of molecular data are available. For example, methylation measurements, protein-protein interactions, and chromosome conformation capture data could be used as additional data layers for predicting gene function. Because many genomic phenomena are tissue-dependent and genes are regulated differently in different tissues and conditions, such as in the above *SLCO* example, it would be useful to incorporate multiple layers of molecular evidence from various tissues to arrive at more accurate and stable gene function predictions.

Third, although we found hundreds of biologically meaningful "transcriptional components" using PCA (chapter 2), there are other possible ways of transforming gene expression data to subcomponents. Independent component analysis (ICA) is a method that finds statistically independent signals that comprise the observed data.[7] As PCA has been useful in identifying biological effects in gene expression data despite the forced orthogonality of principal components, ICA holds the promise of capturing information about the tissue- or condition-dependent biological processes underlying the observed gene expression levels.

Fourth, our method utilised microarray-based gene expression data. Today, tens of thousands of RNA-seq samples are publicly available and can be used to overcome the inherent bias from relying on the probes that are present on the microarray platforms. This is further discussed in section 5.1.3.

5.1.2 DEPICT

To achieve the second aim of this thesis — to suggest which genes and pathways may be relevant to different phenotypes — we devised and developed the DEPICT method, which utilises our gene function predictions in combination with findings from genome-wide association studies. The method has been successfully used to interpret results from several recent GWASs, including studies of human height,[8] body mass index,[9] body fat percentage and distribution,[10,11] and schizophrenia.[12] For example, for human height, DEPICT prioritised ossification and skeletal growth pathways that had not previously been implicated in height. In addition, the prioritised genes were generally highly expressed in relevant tissues such as cartilage, joints and spine. For body mass index, DEPICT prioritised brain-related pathways and found that genes within the associated loci are enriched for expression in the brain and central nervous system, implicating neurological factors play a role in body mass index.

While the basic idea behind DEPICT — that phenotype-relevant genes in the different genetic loci found in a GWAS should share functionality — is not new and has been used in other methods such as GRAIL,[13] it is the gene-pathway scores, calculated systematically from a large, heterogeneous expression dataset, that enable DEPICT to perform better than earlier methods. DEPICT systematically outperformed GRAIL in gene prioritisation when benchmarked against each other in Chron's disease, human height, and lipoprotein cholesterol. DEPICT also predicted more biologically relevant pathways for

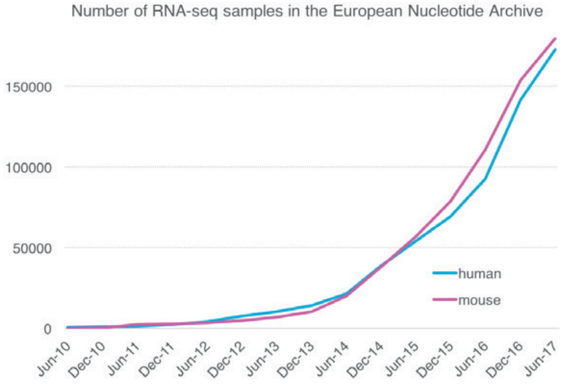


Figure 5.2 *The total number of human and mouse samples in the European Nucleotide Archive. The numbers are growing fast and have nearly doubled each year during the last few years. Data accessed on 18th August, 2017.*

all these phenotypes than MAGENTA,[14] a widely used GWAS gene set enrichment tool.

While the DEPICT method has performed well under the assumption that genes in different loci share functionality, for many complex traits there are several distinct biological processes affecting the phenotype. If some GWAS associated loci mainly affect one biological process or pathway and certain other loci mainly affect another process, and both processes are relevant for the phenotype, the simple functionality-sharing assumption might dilute the findings of our method. One future research opportunity would be to investigate whether, for some complex traits, loci can be grouped according to their downstream consequences using eQTL and co-expression analyses combined with other layers of molecular data. Analysing gene expression, alternative splicing, methylation and protein-protein interactions separately in different tissues that are relevant for a phenotype would help in understanding the context-dependent consequences of genetic variation. Groups of associated loci may then be found that contribute differently to a given phenotype and, by utilising the DEPICT method on such groups of loci, new relevant genes and pathways might be prioritised that would otherwise not be considered significant.

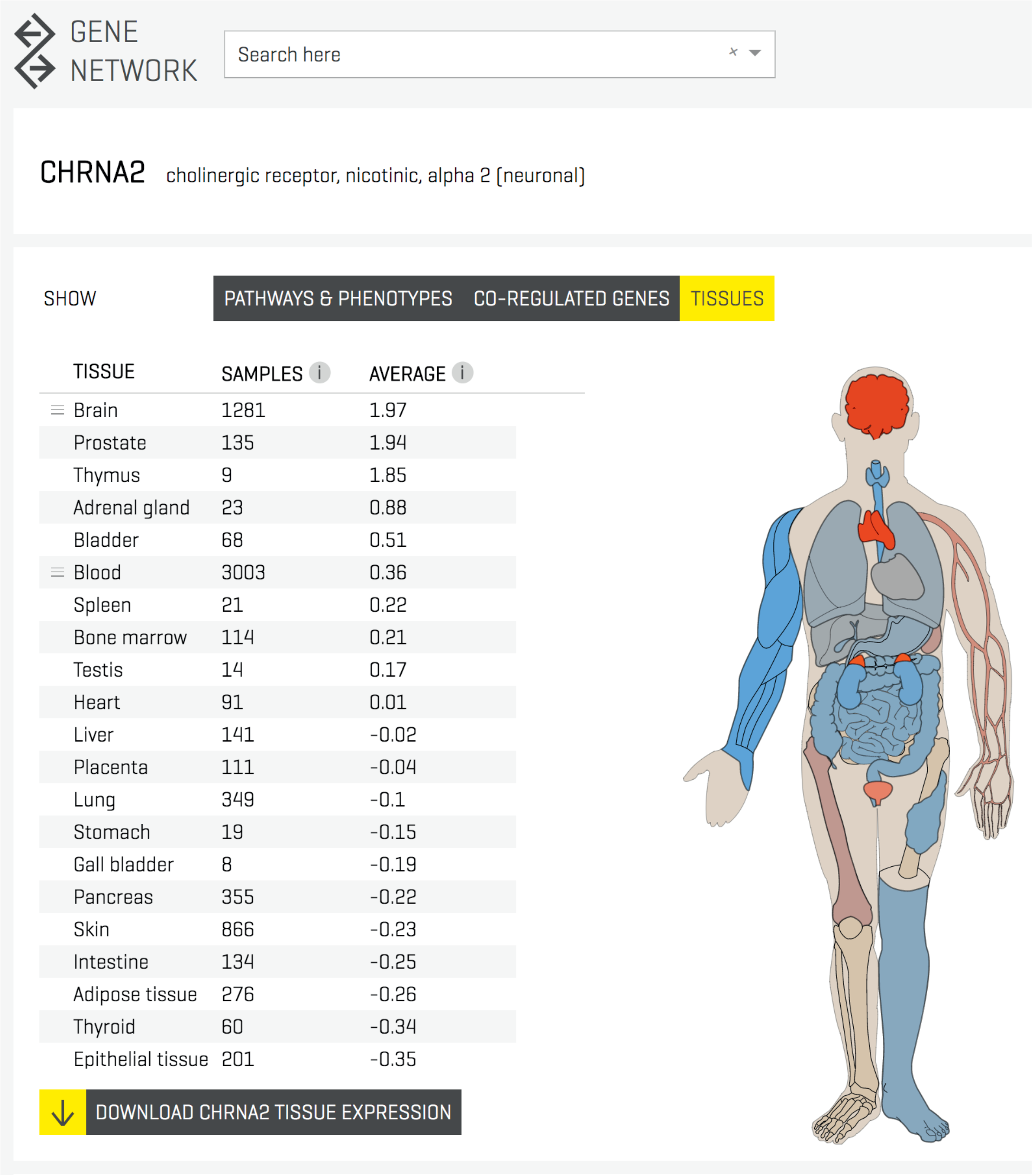
Extending the above line of thinking leads to possibilities for analysis of epistatic effects, i.e. non-linear interactions between several genes or loci. Analyses of epistasis typically suffer from low statistical power,[15] but, if loci can be justifiably grouped according to their downstream consequences, the burden of statistical testing of all pairs of loci can be reduced.

5.1.3 RNA-SEQ

While gene expression levels from microarray experiments have been hugely useful, new technologies to measure gene expression levels in cells have been developed during the last ten years and RNA-seq has become a widely-used method. Microarray studies are inherently limited by the choice of probes on the array, which is based on known genomic information. Microarray measurements also have high background signal levels because of the cross-hybridisation method. In addition, the measured signals can suffer from saturation. RNA-seq technologies do not have these limitations, offering better sensitivity, a higher dynamic range, and truly genome-wide data from being able to measure the whole transcriptome in cells.[16]

In chapter 2, we reanalysed 77,840 high-quality human, mouse and rat microarrays freely available in Gene Expression Omnibus. As of August 2017, the European Nucleotide Archive (ENA) hosts more than 180,000 human RNA-seq samples and more than 190,000 mouse samples, and the numbers continue to grow at a fast rate[17] (see figure 5.2). Although some of these samples have low coverage (i.e. a low number of reads) and some are of suboptimal quality, e.g. because of degraded RNA material, the available sample sizes are such that there is now a lot of potential to be realised in jointly reanalysing data from different RNA-seq experiments.

The data provided by RNA-seq platforms consist of complementary DNA (cDNA) reads. The lengths of the reads range from a few dozen to a few hundred bases. For subsequent genomic analysis, the reads are typically aligned to a reference genome. Aligning the reads in tens or hundreds of thousands of samples can be a time-consuming process even when using computer clusters. To quantify the numbers of transcripts present in the samples in a substantially faster way, a method called “kallisto” has recently been developed.[18] Instead of precisely mapping the RNA-seq reads to a reference, kallisto leverages a pseudoalignment approach to directly map the reads to pre-determined transcripts. This makes the transcript quantification process two orders of magnitude faster than existing alignment-based methods, and large



CHRNA2 cholinergic receptor, nicotinic, alpha 2 (neuronal)

TISSUE	SAMPLES	AVERAGE
Brain	1281	1.97
Prostate	135	1.94
Thymus	9	1.85
Adrenal gland	23	0.88
Bladder	68	0.51
Blood	3003	0.36
Spleen	21	0.22
Bone marrow	114	0.21
Testis	14	0.17
Heart	91	0.01
Liver	141	-0.02
Placenta	111	-0.04
Lung	349	-0.1
Stomach	19	-0.15
Gall bladder	8	-0.19
Pancreas	355	-0.22
Skin	866	-0.23
Intestine	134	-0.25
Adipose tissue	276	-0.26
Thyroid	60	-0.34
Epithelial tissue	201	-0.35

***Figure 5.3** A web interface for investigating the expression of genes in different tissues. Here the expression of CHRNA2 is shown based on 32,000 RNA-seq samples. CHRNA2 has been associated with susceptibility to lung cancer. 20 Human picture courtesy of Thomas Craig and Pytrik Folkertsma. Gene Network logo and look and feel by Clever°Franke.*

numbers of samples can be analysed with fewer computing resources.

To be able to jointly analyse samples from different experiments, the quantified transcript counts have to be jointly normalised. Because many genes or transcripts are expressed in a tissue-specific or cell type-specific manner, and the overall expression is low for many non-coding genes that can regulate gene

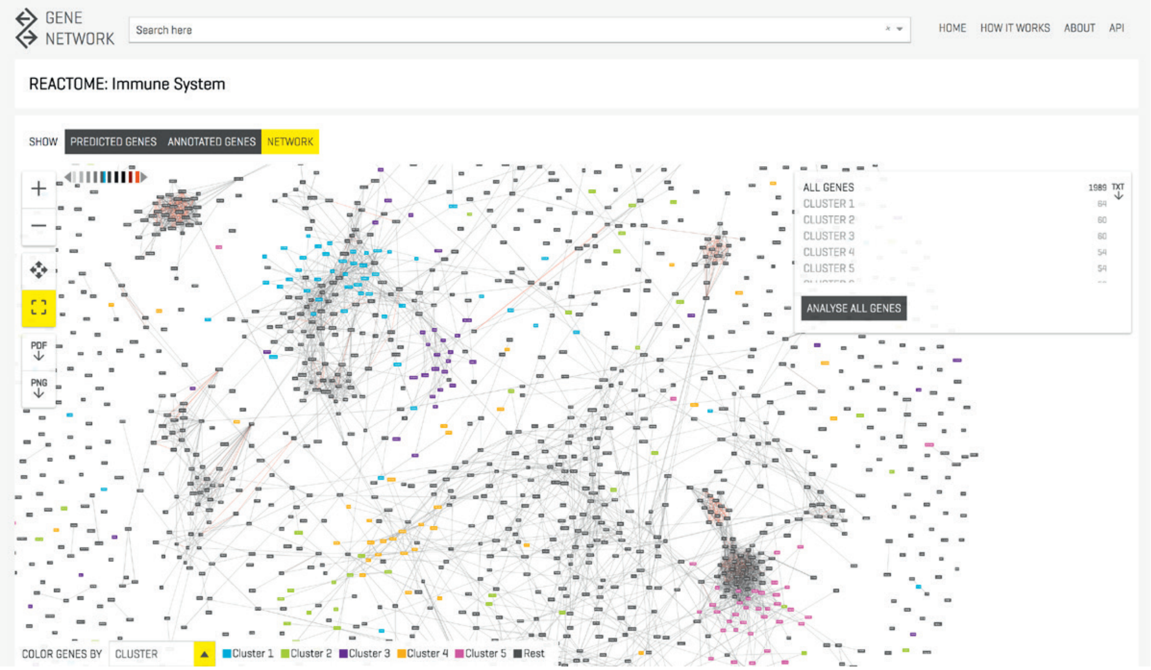


Figure 5.4 *A web interface for investigating gene co-regulation networks and performing pathway analyses. Here a part of the co-regulation network of immune system genes is shown.*

expression in specific tissues, the normalisation process is not trivial. Quantile normalisation is often used for microarray data, but for the sparser RNA-seq data it tends to induce spurious correlations between lower expressed genes. In our experience, the DESeq normalisation method[19] does not suffer from this problem and offers a more robust way of performing normalisation. (Comparison of normalisation methods and implementing the DESeq method, personal communication, Sipko van Dam).

To predict functions for all known human genes, including non-coding ones, we have jointly analysed human RNA-seq samples available in the ENA. We downloaded the raw data for 67,021 thousand samples that had been submitted to ENA before or during the first half of 2016 and that each had at least 500,000 reads. We quantified transcript counts in these samples using kallisto (version 0.42.4). We then applied a stringent quality control procedure, removing samples in which less than 70 % of reads mapped to the transcriptome. Judging by the principal component score distribution of the first principal component, we also removed 4,619 of the remaining samples that were considered to be outliers. In addition, we removed 147 samples that had been run with a non-Illumina platform. For the remaining 31,995 samples that we deemed of high quality, we performed principal component analysis on the gene covariance matrix and identified 307 principal components with a Cronbach's alpha value higher than 0.7. Using these components, we then applied the method described in chapter 2 to construct a gene co-regulation network, predict gene functions, and ascertain the extent of expression for each gene in selected tissues (see figures 5.3 and 5.4).

In comparison with the microarray data that we used in the work described in chapter 2, we observed an increase in the average AUC for Reactome pathways: for microarrays, the average AUC was 0.86, while for the RNA-seq data it was 0.89. However, this difference can be partly explained by the higher number of lower expressed genes in RNA-seq data. Such genes typically have not been assigned to pathways and, because of their low expression, their predictions tend to be less significant than for the more highly expressed genes. This automatically yields a higher AUC value. For a fully justified comparison of prediction accuracy between microarray and RNA-seq data, from the beginning the analyses should be restricted to the genes that are represented both on the microarray platforms and in the RNA-seq data. We investigated the function prediction results based on RNA-seq data for lincRNAs whose expression we had found to be affected by disease-associated SNPs (see chapter 3). The top-predicted Reactome pathways for each

***Table 5.1** The top-predicted Reactome pathways for lincRNAs found to be affected by disease-associated SNPs (see chapter 3). The top-predicted pathways using both microarray data and RNA-seq data are shown.*

associated SNP(s)	eQTL lincRNA	GWAS trait(s)	top-predicted Reactome pathway (microarrays)	p-value (microarrays)	top-predicted Reactome pathway (RNA-seq)	p-value (RNA-seq)
rs13278062	RP11-1149O23.3	Exudative agerelated macular degeneration	Hormone ligandbinding receptors	1.7×10^{-2}	Regulation by c-FLIP	1.1×10^{-10}
rs6490294	MAPKAPK5-AS1	Mean platelet volume	Formation of ATP by chemiosmotic coupling	5.0×10^{-3}	Mitochondrial translation initiation	2.8×10^{-5}
rs10849915	LINC01405	Alcohol consumption	Striated muscle contraction	1.2×10^{-26}	Striated muscle contraction	2.7×10^{-12}
rs7542900	LINC01057	Type 2 diabetes	Trafficking of AMPA receptors	1.5×10^{-2}	Transcriptional regulation of pluripotent stem cells	9.4×10^{-4}
rs17767419, rs3813582	LINC01229	Thyroid volume, thyroid function	N/A	N/A	Eicosanoid ligand binding receptors	2.8×10^{-3}

lincRNA are shown in table 5.1. For the lincRNA that was affected by the SNP associated with age-related macular degeneration, we did not predict any significant pathway based on microarray data. However, with the RNA-seq data, we predicted "regulation by c-FLIP", a relevant pathway for the trait based on functional evidence,[21,22] as the top pathway. In addition, for the lincRNA affected by SNPs associated with thyroid volume and function, we could not previously predict functions because there was no probe for that lincRNA on the microarray platforms. With the RNA-seq data, we predicted "eicosanoid ligand-binding receptors" as the top pathway. Now, it has been shown that thyroxine replacement therapy can ameliorate abnormalities of serum eicosanoid levels.[23] These results demonstrate that jointly analysing large numbers of RNA-seq samples is useful for predicting functions for lincRNAs.

Although gene expression measurements can be of great value for gene predicting function, they only give direct information of the first step in the sequence of events from DNA to organism phenotype. Recent research shows that methods that combine several data sources, co-function networks, and molecular data layers typically perform better than any single prediction approach.[24] In addition, different combinations of data and methods show different performances in different gene function domains such as molecular function and biological process.[25] Thus, to improve the gene function prediction approach described in this thesis, additional data layers, such as protein and methylation measurements, should be used.

5.1.4 CLINICAL APPLICATIONS

In addition to predicting gene functions in the context of pathways as we have described in chapter 2, we also used phenotype information from the Human Phenotype Ontology (HPO) for the RNA-seq data. The HPO provides sets of genes annotated to phenotypic abnormalities (later referred to as simply phenotypes). Using these gene sets, we could predict which as yet unknown genes might also be relevant to various phenotypes. For example, for the Vel blood group culprit, SMIM1, we predicted "Increased red cell osmotic fragility" ($p = 1.6 \times 10^{-10}$) and "Abnormality of the heme biosynthetic pathway" ($p = 1.2 \times 10^{-9}$) as the top phenotypes.

HPO phenotype terms are used in clinics to characterise patients' phenotypes. Since we had comprehensively predicted scores for the likelihood of each human gene affecting each HPO phenotype, this led us to speculate that it should be possible to find previously unknown disease genes causing illness in individual patients (this idea was raised in discussions with Lude Franke and Joeri van der Velde).

We devised a simple method to prioritise genes for an individual patient based on phenotypes that had been assigned to her or him in the clinic. We summed up the prediction Z-scores for each human gene over an individual patient's phenotypes and subsequently ranked the genes based on the resulting total score.

To assess the performance of this method, we prioritised genes for a few critically ill newborns based on their phenotypes (they had each been given a genetic diagnosis). For example, one newborn with flexion contracture and abnormality of muscle morphology was diagnosed, based on whole-genome sequencing, with a homozygous mutation in the *KLHL41* gene (Cleo van Diemen et al., manuscript submitted). The gene ranked 20th out of 56,439 in our prioritisation analysis, based solely on the patient's two phenotypes. *KLHL41* has not been associated to either of these two phenotypes in HPO. Our top-two predicted phenotypes for the gene are "Neck muscle weakness" ($p = 7.1 \times 10^{-18}$) and "EMG abnormality" ($p = 1.1 \times 10^{-16}$). Gene prioritisation results

Table 5.2 *Gene prioritisation results for six critically ill newborns. The prioritisation is based solely on a patient's phenotypes, incorporating no genetic data.*

Human Phenotype Ontology phenotypes	causal gene	rank of the causal gene
leukoencelophalopathy, abnormal central nervous system myelination	EIF2B5	1,029 / 56,439
microcephaly, abnormality of the nervous system	EPG5	1,944 / 56,439
lactic acidosis, abnormality of metabolism / homeostasis, abnormal respiratory system morphology	GFER	2,992 / 56,439
cardiomyopathy, hepatomegaly	GLB1	456 / 56,439
flexion contracture, myopathy	KLHL41	20 / 56,439
global developmental delay, microcephaly, dystonia, hearing impairment, hypomyelination	RMND1	2,308 / 56,439

for all six diagnosed newborns are shown in table 5.2. Further work with a large number of already diagnosed patients with different phenotypes needs to be done to benchmark this approach.

This gene prioritisation method can be made useful for clinical purposes by utilising patients' whole-genome or exome sequencing data. A patient's variants can be classified and prioritised for clinical purposes using bioinformatic methods, such as the recently introduced GAVIN.[26] In combination with variant filtering and prioritisation, the phenotype-based gene prioritisation method offers possibilities to help clinicians reach a genetic diagnosis.

In addition, RNA-seq data from individual patients can be leveraged in diagnostics. Splice variants and exon skipping, for example, can be detected from an RNA-seq sample.[27] One avenue for further research is to investigate whether it is possible to correct individual RNA-seq samples for "healthy variation", to identify aberrantly expressed genes that might be relevant for the patient's disease by using a similar methodology to that described in chapter 2, which we used to detect somatic copy number alterations in cancer samples.

5.2 PAST AND FUTURE PERSPECTIVES

During the last ten years, the technologies used in human genetics have improved tremendously. Nowadays, the sequencing of a whole genome or exome is much faster, more accurate, and less expensive than a decade ago. These technological advances have also led to substantial increases in the sample sizes now being studied. In 2007, GWASs consisted of a few thousand samples,[28] whereas now the biggest studies include hundreds of thousands of individuals.[8] As the cost of genotyping and sequencing continues to decrease, sample sizes are expected to grow towards millions.[29]

In terms of gene expression studies, RNA-seq technologies have superseded microarrays in the last decade. Like GWASs, sample sizes of eQTL studies have increased.[30] Soon, hundreds of thousands of high-quality RNA-seq samples from diverse tissues and conditions will be publicly available for research. As we have learned about the diversity of the workings of genetic variants and genes depending on the tissue, cell type and external stimuli, this large amount of heterogeneous data will be useful in identifying so far unknown key biological processes and the genes driving them.

Despite the usefulness of RNA-seq experiments, "traditional" RNA-seq procedures have, for some research questions, a major drawback: they measure gene expression from bulk populations of cells. Therefore, the resulting gene expression levels reflect the average expression in the cell population. To increase the procedure's resolution to the level of individual cells, single-cell RNA-seq (scRNA-seq) methods have been in development for almost as long as RNA-seq technologies have existed.[31] Although scRNA-seq methods still need to overcome technical artefacts and noise from sample preparation,[32] they can now process hundreds of thousands of cells in parallel, holding the promise of detailed views of the transcriptome in different developmental stages and during tissue differentiation. Novel cell types have already been identified using scRNA-seq.[33]

Bulk RNA-seq experiments have revealed that allele-specific expression (ASE), or imbalance of expression between maternal and paternal alleles, is prevalent in the mouse genome.[34]

Because of the possibility it offers of investigating the early developmental stages of organisms, scRNA-seq can lead us towards a better understanding of ASE and genomic imprinting.

To investigate phenotypic consequences of gene inactivations on the gene expression level, new technologies, such as Perturb-seq, have recently been developed.[35,36] Perturb-seq combines scRNA-seq with CRISPR-mediated perturbations such as gene inactivations. A major novel advantage of these new technologies is that they can be used to study the differences of gene knockout effects from cell to cell. Therefore, they hold the potential of expanding our understanding of gene regulatory networks.

For protein-coding genes, the abundance of RNA does not necessarily directly translate into the amount of protein. Proteomic approaches are needed to accurately determine protein quantities in cells. Protein-protein interaction data can be used to leverage physical contacts between individual proteins. Methylation influences gene expression, and methylation data can explain phenotypic variation to an extent comparable with genetic data.[37] During the last few years, the microbiome composition has been found to influence phenotypes, and the influence of human genetics on the microbiome has been studied extensively.[38] Jointly analysing individuals' genetic data, gene expression, methylation, microbiome composition and lifestyle, and combining this with detailed phenotypic information, such as in the Dutch LifeLines DEEP project,[39] holds potential not for only predicting gene function, but also for use in personalised medicine.

Today, there are major efforts to create large biobanks of both genetic and health record data, such as the population-based Estonian Biobank of the Estonian Genome Centre,[40] or the FinnGen project, which combines data from all the Finnish biobanks.[41] Analysing medical history in combination with genetic data on a large scale holds great promise for personalised medicine. For example, drug responses for diseases that are currently diffcult to medicate correctly, such as depression, might be predicted from an individual's genetic make-up.

In the last few decades, traditional statistical approaches have been extensively used in analysing genetic data. In the last few years, however, sophisticated machine learning (or deep learning) methods have become increasingly popular.[42] These neural network-based methods can perform remarkably well as predictors or classifiers on suffciently large datasets, but their internal workings can be diffcult to interpret. As the amount of data on various molecular levels grows, and deep learning methods develop further, together with our understanding of them, traditional statistics will increasingly give way to these more powerful approaches.

In conclusion, as sample sizes in genetic studies will continue to increase rapidly, we will discover novel genetic variation underlying disease and gain a more comprehensive view of how genetic variation affects phenotypic variation. However, especially for rare diseases, there is a practical limit on the number of samples that can be collected, so the focus of research should not only be on variant hunting, but also on the biological interpretation of the findings. Recent technologies such as scRNA-seq will yield further insights into organism development and tissue differentiation, which will expand our knowledge of disease aetiologies. Reverse genetics methods, such as Perturb-seq, can increase our understanding of the diversity of gene regulation and function. Novel computational methods should be developed to point to not only the statistical relationships in the data, but also the biological processes driving the observations. These methods should incorporate multiple layers of molecular data to be able to yield a more comprehensive picture of genetics in disease and health.

"In coming years, doctors increasingly will be able to cure diseases like Alzheimer's, Parkinson's, diabetes and cancer by attacking their genetic roots", said US President Bill Clinton upon announcing the completion of the draft sequence of the human genome in 2000.[43] Since then this kind of optimism has faded as we have come to better appreciate the true complexity of many complex traits. However, with the current technological advances, we are working in an extremely interesting period of time for human genetics. Because we now know enough to understand how little we know, there are many great opportunities for further research in this wonderful field of science that studies the basis of life!

REFERENCES

[1] A. Cvejic et al. SMIM1 underlies the Vel blood group and influences red blood cell traits.*Nature Genetics*, 45:542–545, Apr 2013.

[2] C. Rietveld et al. GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science*, 340(6139):1467–1471, Jun 2013.

[3] D. Warde-Farley et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38 (Web Server Issue):W214–W220, Jul 2010.

[4] J. Montojo et al. GeneMANIA: Fast gene network construction and function prediction for Cytoscape. *F1000Research*, 3, Jul 2014.

[5] D. Szklarczyk et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45 (Database issue):D362–D368, Jan 2017.

[6] M. Roth, A. Obaidat, and B. Hagenbuch. OATPs, OATs and OCTs: the organic anion and cation transporters of the SLCO and SLC22A gene superfamilies. *British Journal of Pharmacology*, 165(5):1260–1287, Mar 2012.

[7] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

[8] A. Wood et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173–1186, Nov 2014.

[9] A. Locke et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, Feb 2015.

[10] Y. Lu et al. New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature Communications*, 7, Feb 2016.

[11] D. Shungin et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187–196, Feb 2015.

[12] Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511:421–427, Jul 2014.

[13] S. Raychaudhuri et al. Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. *PLOS Genetics*, 5 (6), Jun 2009.

[14] A. V. Segrè et al. Common Inherited Variation in Mitochondrial Genes is not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. *PLOS Genetics*, 6(8), Aug 2012.

[15] S. M. Lorin Crawford and X. Zhou. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genetics*, 13(7), Jul 2017.

[16] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, Jan 2009.

[17] R. Leinonen et al. The European Nucleotide Archive. *Nucleic Acids Research*, 39 (Database Issue):D28–D31, Jan 2011.

[18] N. L. Bray et al. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34:525–527, Apr 2016.

[19] W. H. Michael I Love and S. AndersEmail. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(550), Dec 2014.

[20] J. D. McKay et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature Genetics*, 49:1126–1132, Jun 2017.

[21] Y. Murakami et al. Photoreceptor cell death and rescue in retinal detachment and degenerations. *Progress in Retinal and Eye Research*, 37, Nov 2013.

[22] O. Micheau et al. The Long Form of FLIP Is an Activator of Caspase-8 at the Fas Death-inducing Signaling Complex. *Journal of Biological Chemistry*, 277(47):45162–45171, Nov 2002.

[23] Y. Zhang et al. Thyroxine Therapy Ameliorates Serum Levels of Eicosanoids in Chinese Subclinical Hypothyroidism Patients. *Acta Pharmacologica Sinica*, 37(5):656–663, Mar 2016.

[24] Y. Jiang et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(1), Sep 2016.

[25] B. O. Hansen et al. EnsembleNet: ensemble gene function predictions for Arabidopsis thaliana. *bioRxiv doi: 10.1101/181396*, Sep 2017.

[26] K. J. van der Velde et al. GAVIN: Gene-Aware Variant INterpretation for medical sequencing. *Genome Biology*, 18(6), Jan 2017.

[27] S. A. Byron et al. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics*, 17:257–271, Mar 2016.

[28] T. W. T. C. C. Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, Jun 2007.

[29] P. M. Visscher et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101(1):5–22, Jul 2017.

[30] H.-J. Westra. *Interpreting disease genetics using functional genomics*. 2014.

[31] O. Stegle, S. Teichmann, and J. Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, Mar 2015.

[32] J. K. Kim et al. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications*, 6, Oct 2015.

[33] F. Buettner et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33:155–160, Jan 2015.

[34] J. J. Crowley et al. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics*, 47:353–360, Mar 2015.

[35] A. Dixit et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866, Dec 2016.

[36] B. Adamson et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7):1867–1882, Dec 2016.

[37] S. Shah et al. Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. *American Journal of Human Genetics*, 97(1):75–85, Jul 2015.

[38] M. J. Bonder. *The interplay between genetics, the microbiome, DNA-methylation & gene-expression*. 2014.

[349] Lifelines DEEP. 2017.
https://www.lifelines.nl/researcher/biobank-lifelines/additional-studies/lifelines-deep
[Accessed: 8th Sep 2017].

[40] Estonian Genome Center, University of Tartu. Access to biobank. 2017. http://www. geenivaramu.ee/en/access-biobank
[Accessed: 4th Sep 2017].

[41] Tekes. FinnGen-tutkimus. 2017.
https://extranet.tekes.fi/ibi_apps/WFServlet?IBIF_webapp=/ibi_apps&IBIC_server=EDASERVE&IBIWF_msgviewer=OFF&IBIF_ex=O_PROJEKTI_RAP1&CLICKED_ON=&YPROJEKTI=12808707&YTARKASTELU=Z&YKIELI=S&YHANKETYYPPI=11&IBIAPP_app=opendata&YMUOTO=HTML
[Accessed: 4th Sep 2017].

[42] C. Angermueller et al. Deep learning for computational biology. *Molecular Systems Biology*, 12(7), Jul 2016.

[43] National Institutes of Health — National Human Genome Research Institute. June 2000 White House Event. 2000.
https://www.genome.gov/10001356
[Accessed: 2nd Sep 2017].

6

SUMMARIES

6.1 SUMMARY

Human genetics is an exciting, multidisciplinary field of science. Its focus of investigation is on hereditary diseases in humans, with the ultimate aim of alleviating suffering by treating and preventing various diseases.

The biggest project in biology to date, the Human Genome Project, was completed in 2003 and it unravelled the DNA sequence of the human genome. Since then, the cost and time needed for sequencing and genotyping human genomes have fallen dramatically, which now allows for massive studies to be performed consisting of hundreds of thousands of individuals. In the near future, studies of millions of people are likely to appear.

As a result of the large-scale studies during the past decade, we have come to appreciate the true complexity of genetics. In genome-wide association studies, hundreds of genetic variants have been associated with a specific phenotype or disease. However, these variants together typically only explain a small proportion of the heritability of the phenotype, while the mechanisms by which such variants lead to disease are still mostly unknown.

DNA is a marvellous molecule. It contains information which enables it to replicate itself and make the organism develop in its environment. In cells, parts of DNA (i.e. genes) are transcribed into RNA. Some RNA molecules (protein-coding genes) are further translated into proteins. Other RNAs play different roles, such as regulating the transcription of othergenes.

RNA transcription (gene expression) is the first step in the DNA-to-phenotype process. Therefore, studying abundances of RNA in samples from various tissues can help us to understand the cellular phenomena that lead to disease.

In this thesis, I report on my research using publicly available gene expression data on a large scale to predict the functions of genes and to prioritize genes and pathways that may be relevant to different phenotypes and diseases.

In chapter 1, I present a brief history of genetics describing the major advances that have led us to the current state of the field. I also introduce the aims and content of this thesis.

In chapter 2, I describe the use of gene expression data from 77,840 public microarrays. Using these data in combination with established pathway information, we predicted functions for all the human genes represented on the microarrays. We also used the gene expression data to identify somatic copy number alterations in cancer samples.

In chapter 3, I describe how we found large non-coding RNAs whose expression levels were influenced by genetic variants. We also predicted relevant functions for some of these RNAs whose expression was affected by disease-associated variants.

In chapter 4, I report on how we leveraged the gene function predictions from chapter 2 to prioritize genes and pathways that may be relevant to various phenotypes, based on findings from genome-wide association studies. We developed a computational method called DEPICT that can be used to interpret the results of any genome-wide association study.

In chapter 5, I discuss the findings of this work, as well as its strengths and limitations. I also consider some future directions and clinical applications.

In chapter 6, I present the summary of this work.

6.2 SAMENVATTING

Humane genetica is een spannend en multidisciplinair vakgebied. De focus van onderzoek binnen de humane genetica ligt op erfelijke ziekten, en het doel van het onderzoek is het verlichten van lijden door het behandelen en voorkomen van diverse ziekten.

Het grootste project in de biologie tot dusver, het Humane Genoomproject (the Human Genome Project), is in 2003 afgerond en onthulde voor het eerst de hele DNA-sequentie van het menselijke genoom. Sindsdien zijn de kosten van en tijd die nodig is voor sequencing en genotypering van menselijke genen drastisch gedaald. Dit maakt grote studies van honderden duizenden individuen, en in de nabije toekomst, van miljoenen individuen, mogelijk.

Als gevolg van deze grootschalige genetische studies, realiseren we ons voor het eerst hoe complex de genetica van de mens eigenlijk is. Honderden genetische varianten blijken in genoomwijde associatiestudies met een specifiek fenotype of ziekte geassocieerd. Deze varianten verklaren echter meestal slechts een klein deel van de erfelijkheid van dit fenotype. Bovendien zijn de mechanismen waarmee deze varianten tot ziekte leiden meestal nog steeds onbekend.

DNA is een prachtig en veelzijdig molecuul, het bevat informatie waarmee het repliceert en waarmee het organisme zich in zijn omgeving kan ontwikkelen. In cellen worden delen van het DNA (de genen) overgeschreven naar RNA. Sommige RNA-moleculen (van voor eiwit coderende genen) worden verder omgezet naar eiwitten (translatie). Andere RNA-moleculen spelen uiteenlopende rollen, zoals bijvoorbeeld het reguleren van de transcriptie van andere genen.

RNA transcriptie (ook wel genexpressie genoemd) is de eerste stap in het process van DNA naar fenotype. Daarom kan het bestuderen van kwantiteit aan RNA in monsters uit verschillende weefsels ons helpen bij het begrijpen van cellulaire fenomenen die tot ziekte leiden. In dit proefschrift heb ik grote hoeveelheden publiek beschikbare genexpressiegegevens gebruikt om de functies van genen te voorspellen en genen en de gen-specifieke netwerken (in het Engels: pathways) te prioriteren die relevant kunnen zijn voor verschillende fenotypen en ziekten.

In hoofdstuk 1 presenteer ik een korte geschiedenis van genetica die de vooruitgang tot aan de huidige stand van het veld beschrijft. Ik zet hier ook de doelen van mijn onderzoek en de inhoud van dit proefschrift verder uiteen.

In hoofdstuk 2 gebruikten we genexpressiegegevens van 77.840 microarrays en gepubliceerde gen-netwerk informatie om functies voor alle menselijke genen vertegenwoordigd op de microarrays te voorspellen. We hebben deze genexpressiegegevens ook gebruikt om somatische copynumbervariaties in kankermonsters te identificeren.

In hoofdstuk 3 vonden we grote niet-coderende RNA's, waarvan de expressieniveaus door genetische varianten beïnvloed werden. We hebben ook relevante functies voor sommige van deze RNA's voorspeld waarvan de expressie door ziekteverwante varianten beïnvloed werden. In hoofdstuk 4 hebben we de genfunctievoorspellingen uit hoofdstuk 2 aangewend om genen en pathways te prioriteren die relevant kunnen zijn voor verschillende fenotypen, gebaseerd op bevindingen van genoomwijde associatiestudies. We ontwikkelden een berekeningsmethode genaamd DEPICT die gebruikt kan worden om resultaten van een genoomwijde associatiestudie te interpreteren.

In hoofdstuk 5 heb ik de bevindingen van dit werk besproken, evenals de sterke punten en beperkingen ervan. Ik heb ook een aantal toekomstige richtlijnen en klinische toepassingen overwogen.

In hoofdstuk 6 heb ik dit werk samengevat.

6.3 YHTEENVETO

Ihmisgenetiikka on jännittävä, monitieteinen tutkimusala, joka keskittyy perinnöllisten sairauksien tutkimiseen. Sen perimmäinen määränpää on ihmisten kärsimyksen vähentäminen sairauksia hoitamalla ja ehkäisemällä.

Biologian historian tähän mennessä suurin edesottamus, the Human Genome Project (Ihmisgenomiprojekti), päättyi vuonna 2003. Siinä selvitettiin ihmisen genomin DNA-sekvenssi. Sittemmin ihmisten genomien sekvensoinnin ja genotyypityksen hinta on laskenut huimasti samalla kun niihin tarvittava aika on lyhentynyt. Tämän ansiosta nykyään tehdään massiivisia, satojatuhansia ihmisia käsittäviä tutkimuksia. Lähitulevaisuudessa otosten koot liikkuvat jo miljoonissa.

Kymmenen viime vuoden aikana tehtyjen laajamittaisten tutkimusten tuloksena olemme alkaneet ymmärtää genetiikan monimutkaisuutta aiempaa paremmin. Koko genomin laajuisissa assosiaatiotutkimuksissa on löydetty satoja geneettisiä variantteja, joilla on yhteys tiettyyn fenotyyppiin tai sairauteen. Nämä variantit yhdessä selittävät kuitenkin tyypillisesti vain pienen osan fenotyypin periytyvyydestä. Erityisen huomionarvoista on, että syyt, miksi nämä variantit johtavat sairauden syntymiseen, ovat suurelta osin yhä tuntemattomia.

DNA on verraton molekyyli. Se sisältää tietoa, jonka avulla se monistaa itseään ja saa organismin kehittymään ympäristössään. Soluissa osia DNA:sta (geenit) luetaan RNA:ksi (transkriptio). Jotkin RNA-molekyylit eli proteiineja koodaavat geenit käännetään edelleen proteiineiksi (translaatio). Osalla RNA-molekyyleistä on muita rooleja, kuten toisten geenien transkription säätely.

RNA:n transkriptio eli geenien ilmentyminen on ensimmäinen askel DNA:sta fenotyyppiin johtavassa prosessissa. Siksi RNA:n määrien tutkiminen eri kudoksista otetuista näytteistä voi auttaa ymmärtämään soluprosesseja, jotka johtavat sairauksiin.

Tässä väitöskirjassa hyödynnetään suuria määriä julkisesti saatavilla olevaa geenien ilmentymisdataa. Sen avulla ennustetaan geenien toimintoja sekä priorisoidaan geenejä ja biologisia reittejä, jotka voivat olla merkityksellisiä eri fenotyyppeihin tai sairauksiin liittyen.

Luvussa 1 kerrotaan genetiikan lyhyt historia kuvailemalla kehityskulkua, joka on tuonut meidät tämän kiehtovan tieteenalan nykyiseen vaiheeseen. Samalla esitellään väitöskirjan tavoitteet ja sisältö.

Luvussa 2 hyödynnetään geenien ilmentymisdataa 77 840 mikrosirusta. Käyttämällä sitä yhdessä tunnetun reittitiedon kanssa ennustetaan laskennallisesti toimintoja kaikille ihmisgeeneille, jotka ovat edustettuina näillä mikrosiruilla. Geenien ilmentymisdataa käytetään myös somaattisten kopiolukumuutosten tunnistamiseen syöpänäytteissä.

Luvussa 3 paljastetaan pitkiä ei-koodaavia RNA:ita, joiden ilmentymistasoihin tietyt geneettiset variantit vaikuttavat. Lisäksi ennustetaan oleellisia toimintoja tietyille näistä RNA:ista, joiden ilmentymistasot riippuvat sairauksiin yhdistetyistä varianteista.

Luvussa 4 hyödynnetään luvun 2 geenitoimintoennusteita. Näin pystytään priorisoimaan koko genomin laajuisten assosiaatiotutkimusten tulosten pohjalta geenejä ja reittejä, jotka voivat olla oleellisia eri fenotyypeille. Samalla esitellään laskennallinen menetelmä nimeltään DEPICT, jota voidaan käyttää koko genomin laajuisten assosiaatiotutkimusten tulosten tulkintaan.

Luvussa 5 pohditaan tämän työn tuloksia, vahvuuksia ja rajoituksia. Samassa luvussa esitellään myös joitain tulevaisuuden kehityssuuntia ja kliinisiä sovelluksia.

Luku 6 on yhteenveto tutkimuksesta.

6.4 SAMMANFATTNING

Humangenetik är ett spännande, tvärvetenskapligt vetenskapsområde, vars forskning fokuserar på ärftliga sjukdomar hos människan. Dess yttersta mål är att lindra människors lidande genom att behandla och förhindra olika sjukdomar.

Det största projektet hittills i biologins historia, the Human Genome Project, färdigställdes år 2003. Där fastställdes det mänskliga genomets DNA-sekvens. Sedan dess har kostnaden och tiden som krävs för sekvensering och genotypning av mänskliga genom minskat drastiskt. Detta möjliggör massiva studier som numera kan omfatta hundratusentals individer. I en nära framtid kommer studier bestående av miljoner människor att utföras.

Som ett resultat av de stora studierna under det senaste decenniet har den genetiska komplexiteten börjat förstås allt bättre. I associationsstudier som omfattar genomet i sin helhet (genome-wide association studies), har hundratals genetiska varianter sammankopplats med en specifik fenotyp eller sjukdom. Emellertid representerar dessa varianterna tillsammans typiskt endast en liten del av en fenotyps ärftlighet. Särskilt noterbart är, att de mekanismer genom vilka dessa varianter leder till sjukdom, fortfarande mestadels är okända.

DNA är en underbar molekyl. Den innehåller information genom vilken den replikerar själv och gör att organismen utvecklas i sin miljö. I celler transkriberas delar av DNA (gener) till RNA. Vissa RNA-molekyler (proteinkodande gener) läses vidare till proteiner (translation). Andra RNA-molekyler har andra funktioner, såsom regleringen av transkriptionen av andra gener.

RNA-transkription (genuttryck) är det första steget i DNA-till-fenotyp-processen. Därför kan studier av kvantiteten av RNA i prover från olika vävnader hjälpa oss att förstå cellulära fenomen som leder till sjukdom.

I denna avhandling används omfattande mängder ofentligt tillgänglig genuttryck-data. Med hjälp av dessa kan funktioner av gener förutsägas, samt gener och biologiska reaktionsvägar som kan vara relevanta för olika fenotyper och sjukdomar prioriteras.

I kapitel1 presenteras en kort genetikens historia som beskriver de framsteg somhar lett till det nuvarande läget. Samtidigt introduceras denna avhandlings mål och innehåll.

I kapitel 2 används genuttryck-data från 77 840 mikromatriser. Dessa data används i kombination med fastställd information om reaktionsvägar för att förutsäga funktioner för alla mänskliga gener som är representerade i mikromatriserna. Dessa genuttryck-data används också för att identifiera somatiska kopiantalförändringar i cancerprover.

I kapitel 3 presenteras långa icke-kodande RNA-molekyler, vars uttrycksnivåer påverkades av genetiska varianter. Dessutom förutsägs relevanta funktioner för vissa av dessa RNA-molekyler, vilkas uttryck påverkades av sjukdomsassocierade varianter.

I kapitel 4 används förutsägelser av genfunktioner från kapitel 2 som möjliggör prioritering av gener och reaktionsvägar som kan vara relevanta för olika fenotyper baserat på resultaten av genome-wide association studies. Samtidigt presenteras en beräkningsmetod som kallas DEPICT, vilken kan används för att tolka resultat från genome-wide association studies.

I kapitel5 diskuteras resultaten av denna avhandling, liksom dess starkheter och begränsningar. Dessutom övervägs några möjligheter för framtida utveckling samt kliniska tillämpningar av detta arbete.

I kapitel 6 sammanfattas detta arbete.

6.5 RESUMEN

La genética humana es un campo de ciencia emocionante y multidisciplinario. Su foco de investigación se centra en las enfermedades hereditarias en humanos. El objetivo final es aliviar el sufrimiento mediante el tratamiento y la prevención de diversas enfermedades.

El estudio más importante en biología hasta la fecha, el Proyecto Genoma Humano, fue completado en 2003 y desentrañó la secuencia de ADN del genoma humano. Desde entonces, el costo y el tiempo para secuenciar y genotipificar genomas humanos han disminuido drásticamente, lo que ahora permite estudios masivos compuestos por cientos de miles de personas. En un futuro cercano, es probable que aparezcan estudios de millones de personas.

Como resultado de los estudios a gran escala de la pasada década, hemos aprendido a apreciar la verdadera complejidad de la genética. En estudios de asociación del genoma com-pleto, cientos de variantes genéticas se han asociado con un fenotipo o enfermedad específica. Sin embargo, estas variantes juntas generalmente solo explican una pequeña proporción de la heredabilidad del fenotipo, y los mecanismos por los cuales tales variantes conducen a la enfermedad todavía son en su mayoría desconocidos.

El ADN es una molécula maravillosa. Contiene información que le permite replicarse y hacer que un organismo se desarrolle en su entorno. En las células, partes de ADN (genes) se transcriben en ARN. Algunas moléculas de ARN (genes codificantes) se traducen en proteínas. Otros ARN desempeñan diferentes roles, como la regulación de la transcripción de otros genes.

La transcripción de ARN (expresión génica) es el primer paso en el proceso de ADN a fenotipo. Por lo tanto, estudiar de la abundancia de ARN en muestras de diversos tejidos puede ayudarnos a comprender los fenómenos celulares que conducen a diversas enfermedades.

Para esta tesis, utilicé bases de datos de expresión génica a gran escala disponibles públicamente, para predecir las funciones de los genes, además de priorizar genes y vías que pueden ser relevantes en diferentes fenotipos y enfermedades.

En el capítulo 1, presento una breve historia de la genética describiendo los principales avances que han conducido al estado actual del campo. Además, también presento los objetivos y el contenido de esta tesis.

En el capítulo 2, describo el uso de los datos públicos de expresión génica de 77.840 microarreglos. Usando estos datos en combinación con la información de la vías biológicas establecida, predijimos funciones para todos los genes humanos que fueron representados en los microarreglos. También utilizamos los datos de expresión génica para identificar alteraciones del número de copias somáticas en muestras de cáncer.

En el capítulo 3, describo cómo encontramos ARN largos no codificantes, cuyos niveles de expresión fueron influenciados por variantes genéticas. También predijimos funciones relevantes para algunos de estos ARN, cuya expresión se vio afectada por variantes asociadas con enfermedades.

En el capítulo 4, muestro cómo aprovechamos las predicciones de la función de los genes del capítulo 2 para priorizar los genes y las vías que pueden ser relevantes para diversos fenotipos, basado en los resultados de los estudios de asociación del genoma completo. Desarrollamos un método computacional llamado DEPICT que se puede utilizar para interpretar los resultados de cualquier estudio de asociación del genoma completo.

En el capítulo 5, analizo los resultados de este trabajo, así como sus fortalezas y limitaciones. También considero algunas direcciones futuras y aplicaciones clínicas.

En el capítulo 6 presento un resumen de este trabajo.

6.6 ZUSAMMENFASSUNG

Humangenetik ist ein spannendes und multidisziplinäres Wissenschaftsgebiet. Im Mittelpunkt der Forschung der Humangenetik stehen Erbkrankheiten mit dem Forschungsziel das Leiden durch die Behandlung und Vorbeugung von verschiedenen Krankheiten zu lindern.

Das bisher größte Projekt der Biologie, das Humangenomprojekt (the Human Genome Project), wurde 2003 abgeschlossen und ergab zum ersten Mal die DNA-Sequenz des men-schlichen Genoms. Seitdem sind die Kosten und die Zeit für die Sequenzierung und Genotypisierung des menschlichen Genoms drastisch gesunken. Dies ermöglicht große Studien mit Hunderttausenden von Individuen und in naher Zukunft mit Millionen von Individuen.

Als Ergebnis dieser groß angelegten genetischen Studien stellten wir zum ersten Mal fest, wie komplex die Genetik des Menschen tatsächlich ist. Hunderte von genetischen Varianten sind in genomweiten Assoziationsstudien mit einem spezifischen Phänotyp oder einer bestimmten Krankheit assoziiert. Diese Varianten erklären jedoch meist nur einen kleinen Teil der Vererbung dieses Phänotyps. Wichtig ist, dass die Mechanismen, durch die diese Varianten zu Krankheiten führen, meist noch unbekannt sind.

Die DNA ist ein wunderbares und vielseitiges Molekül. Es enthält Information durch die es sich selbst repliziert und den Organismus in seiner Umgebung entwickeln lässt. In Zellen werden Teile der DNA (die Gene) in RNA transkribiert. Einige RNA-Moleküle (von Protein-kodierende Gene) werden ferner in Proteine übersetzt (die Translation). Andere RNAs spielen andere Rollen, wie zum Beispiel die Regulation der Transkription anderer Gene.

Die RNA-Transkription (Genexpression) ist der erste Schritt im Prozess von der DNA zum Phänotyp. Daher kann das Erforschen der RNA-Menge in Proben aus verschiedenen Geweben helfen, zelluläre Phänomene, die zu Krankheiten führen, zu verstehen.

In dieser Arbeit habe ich große Mengen öfentlich verfügbarer Genexpressionsdaten verwen-det, um Funktionen von Genen vorherzusagen und Gene und Signalwege, die für verschiedene Phänotypen und Krankheiten relevant sein könnten, zu priorisieren.

In Kapitel 1 stellte ich eine kurze Geschichte der Genetik vor die den Fortschritt auf dem aktuellen Stand der Technik beschreibt. Des Weiteren habe ich die Ziele und den Inhalt dieser Arbeit dargelegt.

In Kapitel 2 haben wir Genexpressionsdaten der 77 840 Microarrays ausgewertet. Wir nutzten die Daten in Verknüpfung mit bereits etablierten Signalwegen, um Funktionen aller menschlichen Gene, die auf den Microarrays dargestellt sind, vorherzusagen. Darüber hinaus haben wir diese Genexpressionsdaten zur Identifikation somatischer Kopienzahlvariationen in Krebsproben verwendet.

In Kapitel 3 haben wir lange nicht-kodierende RNAs, deren Expression durch genetische Varianten beeinflusst werden, gefunden. Wir haben zudem relevante Funktionen für einige dieser RNAs, deren

Expression durch krankheitsassoziierte Varianten beeinflusst wurde, prognostiziert.

In Kapitel 4 wurden die Genfunktionsvorhersagen von Kapitel 2 basierend auf Ergebnissen von genomweiten Assoziationsstudien genutzt, um Gene und Signalwege, die für verschiedene Phänotypen relevant sein können, zu priorisieren. Wir haben eine Computermethode namens DEPICT entwickelt, mit der Ergebnisse aus einer genomweiten Assoziationsstudie interpretiert werden können.

In Kapitel 5 habe ich die Ergebnisse dieser Arbeit, sowie ihre Stärken und Grenzen diskutiert und darüber hinaus einige zukünftige Ausrichtungen und klinische Anwendungen betrachtet.

In Kapitel 6 habe ich diese Arbeit zusammengefasst.

7

ABBREVIATIONS

ABBREVIATIONS

ASE	Allele-specific expression	Gene expression levels measured separately for each of the two parental alleles
AUC	Area under the (receiver operating characteristic) curve	The probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance
cDNA	Complementary DNA	DNA synthesised from an RNA template
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats	Bacterial DNA sequences that are useful in genome editing
DEPICT	Data-driven Expression Prioritized Integration for Complex Traits	A method for prioritizing genes, pathways and tissues based on gene expression data, established pathways and GWAS loci
DNA	Deoxyribonucleic acid	A molecule that contains an organism's genetic information
ENA	European Nucleotide Archive	An archive that provides a compherensive record of the world's nucleotide sequencing information
eQTL	Expression quantitative trait locus	A genetic locus that affects RNA expression levels
GAVIN	Gene-Aware Variant INterpretation for medical sequencing	A computational method for classifying genetic variants for clinical diagnostics
GEO	Gene Expression Omnibus	A public functional genomics data repository
GRAIL	Gene Relationships Across Implicated Loci	A computational tool to examine relationships between genes in disease-associated genetic loci
GWAS	Genome-wide association study	An investigation of a group or groups of individuals to find genetic variants associated with a phenotype
HGP	Human Genome Project	An international research project from 1990 to 2003 thatdetermined the sequence of base pairs of human DNA
HPO	Human Phenotype Ontology	A standardized vocabulary of phenotypic abnormalities encountered in human disease
ICA	Independent component analysis	A statistical method for separating a multivariate signal into additive subcomponents
lincRNA	Large/long intergenic non-coding RNA	A transcript longer than 200 nucleotides that doesn't code for a protein
PCA	Principal component analysis	A statistical method for transforming a set of observations into orthogonal components
RNA	Ribonucleic acid	An essential molecule in many cellular processes
RNA-seq	RNA sequencing	A technology to measure the quantity of RNA in a sample
scRNA-seq	Single-cell RNA sequencing	A technology to measure the quantity of RNA in an individual cell
SNP	Single nucleotide polymorphism	A variation in a single nucleotide that occurs at a specific position in the genome
USA	United States of America	

8

ACKNOWLEDGEMENTS

Olisinpa oikea ihminen
olisinpa elävä
olisinpa todellinen
kivun riimut iholla

Olisinpa hetken kerran
elon matkan keston verran
viiman viedä, virran kastaa
käden kovan hyväellä

A. W. Yrjänä — Tuulilukko

The last seven years have been a great many nouns and adjectives for me. Sometimes my focus was on scientific investigation, sometimes on social life, sometimes on self-inquiry, sometimes on nonsense, sometimes on widening of world view, sometimes on exercise. Next to personal growth and all sorts of experiences, I also finished this PhD research in the way I did, in collaboration with many people, and with a reasonable amount of learning.

A lot of people have loved me and cared about me when I have not. I want to professionally and personally acknowledge people who have collaborated with me, supported me, and been awesome during this turbulent inner life time.

Adriaan, thank you so much for all the support, help with practical things, and burritos!

Äiti, kiitos huolenpidosta!

Angelica, it was awesome that you brought some life into the house when you were not work-work-working.

Anna och Niklas, tack så mycket för att ni korrigerade den svenska sammanfattningen!

Annique, your presence among the geeks was welcome. Good discussions about determinism and such.

Anu, kiitos maailman paras sisko!

Arnau, since the Southern attendance in the department decreased, you definitely filled the gap. Never change.

Aska, thank you so much for creating the layout and cover design, submitting the thesis, taking care of the printing, and finishing things off today, on Christmas Eve. In my mind, this was awesome.

Baba, you're the most interesting person that I know. Your story and life experience are extremely inspiring. Here in materially rich countries we are a little bit lost. I see my own pretentiousness better in your company. Principle!

Barbara, thank you for the physio help!

Bart, thank you for teaching me Spanish and all the beers!

Betti, thank you for your sisterly advice and support as well as silly times and yoga teaching!

Casper, you're an amazing friend. Thank you for the excellent discussions and long nights!

Cleo, thank you for all the fun times! Never forget the ninja.

Cleo, thank you for showing me the clinical world a bit! Too bad we never got to combine things more. The tissue prioritization for sample acquisition would've been awesome!

Cisca, I'm grateful for your unshakable belief in me. Despite all my troubles, it was great to work in a competent, well-run department. Your broad view of science is inspiring. I still don't unconditionally believe in inpedendent causal SNPs.

Corien, thank you so much for your long-standing help. Cold comfort for change. As a detail, I enjoyed our interpretation of dreams and such.

Danny, thanks for the interesting discussions. "Self-confidence is like money: it doesn't exist." It was great that you put your LaTeX thesis in GitHub. It saved a lot of time for me.

Dasha, it was nice to go climbing together. I learned a lot of the sport from you. You're a very sweet person.

Dennis, thank you for suggesting technologies such as React and help with the API definitions!

Essi, kiitos kannustuksesta!

Fiaz, ya mon. You've been a great traveling companion. Let's get our minds off our heads more.

Freerk, you're an awesome person. I hope we can find a new beer for you in the party.

Gabriela, oye! Your love, while not always well received, was the best thing that happened to me. I'm very grateful for all the support.

Genaro, thanks for all the fun times and swimming! Estamos chupando tranquilos.

Gosia, I learned a lot from you as you're very strong and fight in the face of bad attitude. Full speed. And thanks for the hangover beer!

Harm-Jan, I couldn't have wished for a better fellow PhD student next to me. Thank you for the support and all the fun things such as rearranging keys.

Heather, you're the best. Thank you so much for caring and helping me through some of the most difficult days of my life. That ice cream was awesome.

Hedi, schönen Dank für Deine Hilfe bei der Übersetzung! Und für die Umarmungen!

Heikki, kiitoksia johdantoni kommentoimisesta, tyyssijasta sekä ~~vaihtelevan tasoisista~~ ah, niin ratkiriemukkaista puujalkavitseistä! Annat maailmaa poikkeuksellisen paljon nähneenä ihmisenä esimerkkiä huumorin merkityksestä.

Helena, isot kiitokset kansitaiteesta! Ja supliikista! Ja.. no.

Inga, kannatti jäädä bisselle!

Isis, what joy! Thank you for all the great times and cheers and beers and showing around in Mexico!

Isä, kiitoksia ymmärryksestä!

Jackie, thank you for listening and wisdom!

Jackie, thank you for the editing and language tips and super detailed considerations! That means to say, from which I learned a lot!

Jan Willem, I only saw you a couple of times but you did and said the best things at the right time. Your help was invaluable.

Javier, thank you for the fun times and reflection!

Jihane, your decency was important among the wild.

Jing, your cheerfulness was amazing. Sometimes it's the small things: thank you for sending me that job posting!

Joeri, could be worse! You brought HPO ideas and clinical viewpoints and affinity propagation and you're awesome.

Johan, you are rather convincing. Your ability to listen, and your carefully but confidently stated insights and directions have been of enormous help to me.

Juti, meigäläinen teigäläinen... tattis haljasista separoinneista ja huikeista keskusteluista. Attic Society forever.

Kieu, your smile brings a cheer!

Kim, your positivity and seeming immunity to adverse situations were astonishing. Your pragmatic thinking helped me a lot, from work to relationships.

Lasse, kiitos että olet isoveljeni!

Lassi, kiva kun tulit käymään!

Lude, thank you for supervising my PhD research. I learned a lot about science from you, from statistical genetics via generation of hypotheses to presentation of results. I enjoyed our — to use a word we both love — brainstorming sessions, in which it sometimes seemed there was only one mind in the room. I also learned a lot from the later ego fights. I believe you also influenced the discovery of my talent in visualisation and interface design. My fondest memory is from the ESHG conference in Amsterdam where we inhaled helium from party balloons. I apologize for the poem and other insults.

Luzma, thank you for nice company and showing around in Spain!

Mamen, you're a very warm person. Don't worry about the English: you speak more important languages.

Manski, olen hyvin kiitollinen isoveljellisestä avunannosta ja kaikenlaisista tykittelyistä vuosien varrella. Korviini on kantautunut uutta tietoa äijän nimen etymologiasta. Kuten tässä vaiheessa hyvin tiedetään — oli kyse g) "en, etiikasta" tai e) "ty, molo", g) I/O:ista — ilmeisin selitys ei aina ole se oikea. Koko tarina jää myöhemmin kerrottavaksi, mutta kuulinpa hiljan (en siis muinaisina aikoina) Ansbachissa kysyttävän, onko ukkeli

a) M. A. ns. Numminen,

b) M. A. ns. Alanko vai

c) M. A. ns. Kimmo Kinnunen.

Manu, kiitoksia saunoista ja nauruista ja harrastushommista.

Marc Jan, the nicest Dutch man on earth volgens mij.

Maria, you are awesome and a great cheer for the deparment. Thank you for your emotional support. Don't let them wear your awesome smile off you. Muchas gracias por tu ayuda con el resumen!

Marijn, thank you for your support when I needed it!

Mitja, you're a great friend and I've spent some of the most memorable times with you. Mountains, lakes, spritzers, barbecue, parties, social philosophy, sailing, psychedelia and Aki Kaurismäki. Who else could I have all this with? Beloved is the one who sits down.

Niek, you're very cool. And you understand hitchhiking.

Noortje, you're an amazing person. I will always remember your angelic help during a sad moment. Also, I can't thank you enough for establishing contact across the ocean!

Olga, thank you for good company!

Patrick, you're brilliant and strong in the face of adversity. I wish I learned more of your pragmatic thinking.

Pieter, thank you for all the computational help, geeky discussions, and whatnot!

Pomme, kiitos kämpästä, tuli tarpeeseen!

Pytrik, you're great! Your capacity to learn combined with your design skills yielded very cool things!

Rob, thank you for providing means for excellent exercise for my earthly temple-vehicle.. where pleasure meets pain. Sorry for forgetting my automated class enrollment program running when going for vacation. I love burpees.

Rodrigo, you smooth motherfucker. Fun times.. hähhähhää! Number six!

Rudolf, I learned statistics and good thinking from you!

Saana, olen hyvin kiitollinen tuesta ja ymmärryksestä.

Sampsa. "Suomalainen ajattelee, että muut ovat häntä huonompia. Hollantilainen ajattelee, että hän on muita parempi." on eräs huikeimmista huomioista ihmiskunnan historiassa. Toivottavasti päästään taas pian seuraamaan muurahaisten edesottamuksia Auringon paistaessa parahultaisesti.

Sasha, thank you for wisdom, climbing company and D'. You were right about the alphabetical order of genes but I'm stubborn.

Sebo, thank you for believing in me and Gene Network fandom!

Taru, kiitoksia jargonin suomentamisesta!

Tessa, you did some of the most important Gene Networkwork. I shouldn't have asked you to do something and proceed to do it myself anyway. Definitely learned from that.

Thomas, "mihin ikinä mä menin...", kiitoksia veljellisistä meiningeistä.

Timo, kiitos yösijasta, hyvästä seurasta ja kaikesta hauskasta!

Tõnu, your dedication to science is astonishing. Thank you for inspiration!

Tune, thank you for making our work visible!

Urmo, you brought proper Nordic (Northern?) spirit in. Also, thank you for showing around Tartu!

Vesa ja Terhi, kiitos että olin aina tervetullut! Olette parhaita.

Vinod, thank you for lincRNA adventures, badminton and inspiration! You're super cool.

Wille, kiitos yksiinvetäisyn finskaamisesta! Aina oppii kielestä näin muodoin.

Wouter, people have often acknowledged me for philosophical discussions. I find this slightly strange because I didn't have a proper philosophical discussion before you entered.

Yang, thank you for bringing smile and for the badminton company. I wish we could do it more.

Zuzanna, thank you for coming to see me, twice! That meant a lot. Also, thanks for the vodka...

Viimeiseksi vaan ei vähäisimmäksi, minä ja järkeni kiittää säilymisestään kaikkia kun kävi visiitillä kaukaisessa Groningenissa: Anu, Eemil, Hedi, Heikki, Heikki, Iivari, JP, JP, Janski, Jaska, Johanna, Jokke, Jude, Jussi, Juti, Lassi, Leila, Lemmy, Lilli, Linda, Manski, Maria, Matti, Merirosvo, Mikko, Mokki, Osku, Pachva, Puti, Riikka, Ripa, Saana, Salama, Samppi, Santtu, Terhi, Thomas, Timo, Vekkuli, Vesa, Ville ja Wille.

Last, thank you everyone who provided me with a place to rest my head during my travels. You know who you are.

Lopuksi kiitän Ismo Alankoa jatkuvasta tuesta, inspiraatiosta, neuvonannosta sekä erinomaisista vaikutteista.

9

LIST OF PUBLICATIONS

2013

V. Kumar, H.-J. Westra, J. Karjalainen, D. V. Zhernakova, T. Esko, B. Hrdlickova,R. Almeida, A. Zhernakova, E. Reinmaa, U. Võsa, et al. Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. *PLoS Genetics*, 9(1):e1003201, 2013

A. Cvejic, L. Haer-Wigman, J. C. Stephens, M. Kostadima, P. A. Smethurst, M. Frontini, E. van den Akker, P. Bertone, E. Bielczyk-MaczyÒska, S. Farrow, et al. SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nature Genetics*, 45(5):542–545, 2013

C. A. Rietveld, S. E. Medland, J. Derringer, J. Yang, T. Esko, N. W. Martin, H.-J. Westra, K. Shakhbazov, A. Abdellaoui, A. Agrawal, et al. GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment. *Science*, 340(6139):1467–1471, 2013

H.-J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, 2013

2014

S. van Sommeren, M. Janse, J. Karjalainen, R. Fehrmann, L. Franke, J. Fu, and R. K. Weersma. Extraintestinal Manifestations and Complications in Inflammatory Bowel Disease: From Shared Genetics to Shared Biological Pathways. *Inflammatory Bowel Diseases*, 20(6):987–994, 2014

Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014

F. Geller, B. Feenstra, L. Carstensen, T. H. Pers, I. A. van Rooij, I. B. Körberg, S. Choudhry, J. M. Karjalainen, T. H. Schnack, M. V. Hollegaard, et al. Genome-wide association analyses identify variants in developmental genes associated with hypospadias. *Nature Genetics*, 46(9):957–963, 2014

V. Kumar, S.-C. Cheng, M. D. Johnson, S. P. Smeekens, A. Wojtowicz, E. Giamarellos-Bourboulis, J. Karjalainen, L. Franke, S. Withoff, T. S. Plantinga, et al. Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia. *Nature Communications*, 5, 2014

C. A. Rietveld, T. Esko, G. Davies, T. H. Pers, P. Turley, B. Benyamin, C. F. Chabris, V. Emilsson, A. D. Johnson, J. J. Lee, et al. Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences*, 111(38):13790–13794, 2014

B. Hrdlickova, V. Kumar, K. Kanduri, D. V. Zhernakova, S. Tripathi, J. Karjalainen, R. J. Lund, Y. Li, U. Ullah, R. Modderman, et al. Expression profiles of long non-coding RNAs located in autoimmune disease-associated regions reveal immune cell-type specificity. *Genome Medicine*, 6(10):1, 2014

A. R. Wood, T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, A. Y. Chu, K. Estrada, J. Luan, Z. Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173–1186, 2014

2015

C.-A. Brandsma, M. van den Berge, D. S. Postma, M. R. Jonker, S. Brouwer, P. D. Paré, D. D. Sin, Y. Bossé, M. Laviolette, J. Karjalainen, et al. A large lung gene expression study identifying fibulin-5 as a novel player in tissue repair in COPD. *Thorax*, 70(1):21–32, 2015

R. S. Fehrmann, J. M. Karjalainen, M. Krajewska, H.-J. Westra, D. Maloney, A. Simeonov, T. H. Pers, J. N. Hirschhorn, R. C. Jansen, E. A. Schultes, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics*, 47(2): 115–125, 2015

V. Kumar, J. Gutierrez-Achury, K. Kanduri, R. Almeida, B. Hrdlickova, D. V. Zhernakova, H.-J. Westra, J. Karjalainen, I. Ricaño-Ponce, Y. Li, et al. Systematic annotation of celiac disease loci refines pathological pathways and suggests a genetic explanation for increased interferon-gamma levels. *Human Molecular Genetics*, 24(2):397–409, 2015

T. H. Pers, J. M. Karjalainen, Y. Chan, H.-J. Westra, A. R. Wood, J. Yang, J. C. Lui, S. Vedantam, S. Gustafsson, T. Esko, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, 6, 2015

W. Peyrot, S. Lee, Y. Milaneschi, A. Abdellaoui, E. Byrne, T. Esko, E. de Geus, G. Hemani, J. Hottenga, S. Kloiber, et al. The association between lower educational attainment and depression owing to shared genetic

effects? Results in ~25 000 subjects. *Molecular Psychiatry*, 20(6):735–743, 2015

D. Shungin, T. W. Winkler, D. C. Croteau-Chonka, T. Ferreira, A. E. Locke, R. Mägi, R. J. Strawbridge, T. H. Pers, K. Fischer, A. E. Justice, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187–196, 2015

P. Deelen, D. V. Zhernakova, M. de Haan, M. van der Sijde, M. J. Bonder, J. Karjalainen, K. J. van der Velde, K. M. Abbott, J. Fu, C. Wijmenga, et al. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Medicine*, 7(1):30, 2015

A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538): 197–206, 2015

M. C. Cornelis, E. M. Byrne, T. Esko, M. A. Nalls, A. Ganna, N. Paynter, K. L. Monda, N. Amin, K. Fischer, F. Renstrom, et al. Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Molecular Psychiatry*, 20(5):647–656, 2015

H.-J. Westra, D. Arends, T. Esko, M. J. Peters, C. Schurmann, K. Schramm, J. Kettunen, H. Yaghootkar, B. P. Fairfax, A. K. Andiappan, et al. Cell specific eQTL analysis without sorting cells. *PLoS Genetics*, 11(5):e1005223, 2015

I. Surakka, M. Horikoshi, R. Mägi, A.-P. Sarin, A. Mahajan, V. Lagou, L. Marullo, T. Ferreira, B. Miraglio, S. Timonen, et al. The impact of low-frequency and rare variants on lipid levels. *Nature Genetics*, 47(6):589–597, 2015

T. W. Winkler, A. E. Justice, M. Graff, L. Barata, M. F. Feitosa, S. Chu, J. Czajkowski, T. Esko, T. Fall, T. O. Kilpeläinen, et al. The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. *PLoS Genetics*, 11(10):e1005378, 2015

2016

C. Pattaro, A. Teumer, M. Gorski, A. Y. Chu, M. Li, V. Mijatovic, M. Garnaas, A. Tin, R. Sorice, Y. Li, et al. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nature Communications*, 7, 2016

C. Ibrahim-Verbaas, J. Bressler, S. Debette, M. Schuur, A. Smith, J. Bis, G. Davies, S. Trompet, J. Smith, C. Wolf, et al. GWAS for executive function and processing speed suggests involvement of the CADM2 gene. *Molecular Psychiatry*, 21(2):189–197, 2016

Y. Lu, F. R. Day, S. Gustafsson, M. L. Buchkovich, J. Na, V. Bataille, D. L. Cousminer, Z. Dastani, A. W. Drong, T. Esko, et al. New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature Communications*, 7, 2016

I. Ricaño-Ponce, D. V. Zhernakova, P. Deelen, O. Luo, X. Li, A. Isaacs, J. Karjalainen, J. Di Tommaso, Z. A. Borek, M. M. Zorro, et al. Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs. *Journal of Autoimmunity*, 68:62–74, 2016

N. Amin, K. V. Allebrandt, A. van der Spek, B. Müller-Myhsok, K. Hek, M. Teder-Laving, C. Hayward, T. Esko, J. G. van Mill, H. Mbarek, et al. Genetic variants in RBFOX3 are associated with sleep latency. *European Journal of Human Genetics*, 24(10):1488–95, 2016

2017

C. R. Marshall, D. P. Howrigan, D. Merico, B. Thiruvahindrapuram, W. Wu, D. S. Greer, D. Antaki, A. Shetty, P. A. Holmans, D. Pinto, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*, 49(1):27–35, 2017

E. A. Nibbeling, A. Duarri, C. C. Verschuuren-Bemelmans, M. R. Fokkens, J. M. Karjalainen, C. J. Smeets, J. J. de Boer-Bergsma, G. van der Vries, D. Dooijes, G. B. Bampi, et al. Exome sequencing and network analysis identifies shared mechanisms underlying spinocerebellar ataxia. *Brain*, 140(11):2860–2878, 2017