UNIVERSITY OF CALIFORNIA, SAN DIEGO

Methods for Identifying Pathway Epistasis and the Effects of Aging on the Human Methylome

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Bioengineering

by

Gregory John Hannum

Committee in charge:

Professor Trey Ideker, Chair
Professor Xiaohua Huang
Professor Richard Karp
Professor Dan O'connor
Professor Kang Zhang

2012

The Dissertation of Gregory John Hannum is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2012

DEDICATION

To my wife Alicia,
who has stood by me in all things,
and shared her patience, creativity, and dedication,
to achievements which are only possible
with a clear vision of what we might become.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Trey Ideker for his support of creativity, critical insights, professional writing, and an unwavering belief that his students can achieve the highest of excellence.

I would like to acknowledge Rohith Srivas for his great collaborative work that resulted in two co-authorships.

I would like to acknowledge Justin Guinney at Sage Bionetworks for his support on the aging project, demonstrating that teamwork and technology can make long-distance collaborations highly effective.

I would like to acknowledge Professor Kang Zhang for his relentless energy and confidence in my abilities.

Chapter 1, in full, is a reprint of the material as it appears in PLoS Genetics, 2009. Rohith Srivas, Aude Guenole, Haico von Attikum, Nevan Krogan, Richard Karp, and Trey Ideker were co-authors. The dissertation author and Rohith were the primary investigators and authors of this paper.

Chapter 2, in full, is a reprint of the material as it appears in Nature Protocols, 2011. Rohith Srivas, Johannes Ruscheinski, Keiichoro Ono, Peng-Liang Wang, Michael Smoot, and Trey Ideker. The dissertation authors and Rohith were the primary investigators and authors of this paper.

Chapter 3, in full, is a reprint of the material as it was submitted to Nature, 2012. Justin Guinney, Li Zhang, Ling Zhao, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Indika Rajapakse,

Stephen Friend, Trey Ideker, and Kang Zhang were co-authors on the paper. The dissertation author and Justin were the primary investigators and authors of this paper.

VITA

2002 Bachelor of Science, University of California, San Diego

2007 Master of Science, University of California, San Diego

2012 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

"Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox" Nature Protocols, 2007.

"Genome-Wide Association Data Reveal a Global Map of Genetic Interactions among Protein Complexes" PLoS Genetics, 2009.

"Assembling Global Maps of Cellular Function through Integrative Analysis of Physical and Genetic Networks" Nature Protocols, 2011.

"Hair Cortisol Level as a Biomarker for Altered Hypothalamic-Pituitary-Adrenal Activity in Female Adolescents with Posttraumatic Stress Disorder After the 2008 Wenchuan Earthquake" Biological Psychiatry, 2011.

"High temperature requirement factor A1 (HTRA1) gene regulates angiogenesis through transforming growth factor-β family member growth differentiation factor 6." Journal of Biological Chemistry, 2012.

"Genome-wide Methylation Profiles Reveal Quantitative Views of Human Ageing Rates" In submission at Nature, 2012

FIELDS OF STUDY

Major Field: Bioengineering

Studies in Bioinformatics and Genomics
Professor Trey Ideker

ABSTRACT OF THE DISSERTATION

Methods for Identifying Pathway Epistasis
and the Effects of Aging on the Human Methylome

by

Gregory John Hannum

Doctor of Philosophy of Bioengineering

University of California, San Diego, 2012

Professor Trey Ideker, Chair

Large-scale molecular data has revolutionized the field of biology. However, such data come with considerable challenges in experimental design, computational modeling, and interpretation. Here I present three papers which take advantage of large-scale molecular data and advance the statistical methods for identifying biologically relevant features. The first paper relates epistatic interactions identified

using gene knockouts to those identified using naturally occurring genetic variants. These interactions are further integrated with physical interaction data to produce a functional cellular map of yeast. The second paper implements genetic and physical interaction alignment in a user-friendly and interactive software package. The third paper investigates the effect of aging on the human methylome. Together, these works represent my principal contribution to the biological community to date and form the basis of my dissertation.

INTRODUCTION

The last decade has been an extraordinary time for the field of molecular biology. The development of high-throughput approaches for molecular screening has allowed researchers to measure the levels of thousands to millions of molecular markers in a single sample. Specifically, the advancement of array profiling and DNA sequencing methods have produced system-wide measurements of the genome and RNA transcription for hundreds of thousands of samples across dozens of organisms. In addition, equally impressive technology for the measurement of molecular interaction networks has yielded maps of the relationships between genes and proteins.

These approaches have been further supported by the silicon revolution. The remarkable pace of improving computational power, algorithm design, and statistical methods have contributed to an environment where whole cellular systems are routinely investigated for novel findings. In addition, communities of researchers have constructed knowledge databases, collating literature on genomic variants, genes, proteins, metabolites, and phenotypes. While far from complete, these databases provide a framework for interpreting new findings and aggregating them in an iterative fashion. These integrative strategies promise a bright future, with the most relevant biological information easily accessible to scientists, doctors, and patients. Profound discoveries underlying disease and treatments should also follow.

Such large data sets come with a number of obstacles that stand in the way of appropriate use and interpretation. Among the biggest obstacles are the problem of statistical noise and power. By screening thousands of molecular profiles for a specific

pattern (ex. differences between cases and controls), statistical noise will produce many strong associations even in the absence of a true biological effect. In addition, even slight biases due to confounding variables or improper model assumptions can create the illusion of meaningful results, which often waste great resources in attempts at validation. For several years these issues plagued the high-throughput community, generating considerable skepticism from the lack of reproducibility in high-profile results. Scientific journals and reviewers quickly became much more careful about enforcing rigorous experimental design and statistical analysis, moving systems biology into the forefront of biological research.

One approach to dealing with statistical noise in high-throughput data is to integrate multiple sources of data to generate a more robust model. This is particularly useful in the model yeast organism *Saccharomyces Cerevisiae*, as it has a great deal of diverse whole-genome measurements. An important example is the network of protein-protein interactions, which characterize the physical binding of proteins. At the same time, high-throughput techniques have been developed to measure genetic, or epistatic, interactions, which represent the cooperative functional effects of gene pairs (i.e. whether two genes independently or cooperatively contribute to a given phenotype). Both of these networks are subject to very high false-negative and false-positive rates. However, when combined together they reinforce strengths and complement weakness, producing relatively accurate cellular maps of functional links between protein complexes and pathways.

Another type of genome-wide data consists of measurements of genetic variations and their correlates to gene expression, known as expression quantitative trait loci (eQTLs). We proposed that eQTLs can be investigated for genetic interactions and integrated with known physical complexes in a similar manner to traditional genetic interaction networks. Our work in this topic was published in PLoS Genetics under the title "Genome-wide association data reveal a global map of genetic interactions among protein complexes." This work is reproduced as Chapter 1.

Such integrative techniques are widely applicable to many data sets, though their implementation can take considerable time. Furthermore, it is very useful to have interactive visualizations of the results. To address these needs, we developed a plugin for the network visualization tool Cytoscape which integrates physical and genetic interaction networks, and produces a detailed interactive cellular map. This work was published in Nature Protocols under the title "Assembling Global Maps of Cellular Function through Integrative Analysis of Physical and Genetic Networks". This work is reproduced as Chapter 2.

Differences between model organisms and humans make it challenging to generate genome-wide network data. Large-scale data for humans is primarily based on DNA sequencing, including the measurement of genomic variants and transcription levels. Recently, genome-wide DNA methylation profiling has become affordable, driving a wave of new findings tying methylation to molecular profiles and diseases such as cancer and cardiovascular disease. One of the strongest associations is the effect of age on the methylome. We used this premise to generate a model of the aging

methylome and to use it to demonstrate differences in the apparent rate of aging between men and women. This work is currently in review at Nature under the title "Genome-wide Methylation Profiles Reveal Quantitative Views of Human Ageing Rates". This work is reproduced as Chapter 3.

Supplemental materials for the published works can be found online at their respective journals.

CHAPTER 1: GENOME-WIDE ASSOCIATION DATA REVEAL A GLOBAL MAP OF GENETIC INTERACTIONS AMONG PROTEIN COMPLEXES

Chapter 1.1: Abstract

This work demonstrates how gene association studies can be analyzed to map a global landscape of genetic interactions among protein complexes and pathways. Despite the immense potential of gene association studies, they have been challenging to analyze because most traits are complex, involving the combined effect of mutations at many different genes. Due to lack of statistical power, only the strongest single markers are typically identified. Here, we present an integrative approach that greatly increases power through marker clustering and projection of marker interactions within and across protein complexes. Applied to a recent gene association study in yeast, this approach identifies 2,023 genetic interactions which map to 208 functional interactions among protein complexes. We show that such interactions are analogous to interactions derived through reverse genetic screens and that they provide coverage in areas not yet tested by reverse genetic analysis. This work has the potential to transform gene association studies, by elevating the analysis from the level of individual markers to global maps of genetic interactions. As proof of principle, we use synthetic genetic screens to confirm numerous novel genetic interactions for the INO80 chromatin remodeling complex.

Chapter 1.2: Author Summary

One of the most important problems in biology and medicine is to identify the genetic mutations that affect human traits such as blood pressure, longevity, and onset

of disease. Currently, large scientific teams are examining the genomes of thousands of people in an attempt to find mutations present only in individuals with certain traits. Until now, mutations have been largely examined in isolation, without regard to how they work together inside the cell. However, large pathway maps are now available which describe in detail the network of genes and proteins that underlies cell function. Here we show how to take advantage of these pathway maps to better identify relevant mutations and to show how these mutations work mechanistically. This basic approach of combining genetic information with known maps of the cell will have wide-ranging applications in understanding and treating disease.

Chapter 1.3: Introduction

A central challenge in genetics is to understand how interactions among different genetic loci contribute to complex traits[1–7]. In model organisms such as yeast, genetic interactions are typically identified using reverse genetic approaches, in which different pairs of genes are systematically knocked out to create a collection of double mutants. Genetic interaction is indicated when the growth rate of the double mutant is slower than expected (e.g., synthetic sickness or lethality) or faster than expected (e.g., suppression)[4,8,9]. Rapid screening of such interactions has been made possible through a variety of methods including Synthetic Genetic Array (SGA) analysis[4], diploid Synthetic Lethality Analysis by Microarray (dSLAM)[3], and epistatic miniarray profiles (E-MAP)[1,2,10,11].

In higher eukaryotes such as humans, reverse genetic analysis has not been so straightforward. Complex traits such as body weight or disease onset can be difficult to study in a cell-based assay, and null mutations are expensive to induce in mammals[12]. Instead, interactions amongst loci have been largely mapped from data generated through forward genetic approaches, such as genome-wide linkage[13] or genome-wide association studies (GWAS)[14,15]. Such methods leverage naturally occurring mutations in the genome to pinpoint loci that have associations, ideally causal associations, with a trait of interest[7].

Mapping pair-wise locus associations has proven remarkably difficult, however. The most basic approach is to perform an exhaustive two-dimensional (2D) scan, in which all pairs of genetic markers are tested for joint association with the phenotype. Because billions of marker pairs must be tested, 2D scans are computationally demanding and suffer from low statistical power due to multiple hypothesis testing. One method to partially mediate this problem is to initiate searches for pair-wise interactions only for markers with strong individual effects[14,15]. Two recent studies by Storey *et al.* and Litvin *et al.* used this approach while accounting for information shared across multiple traits to further enhance statistical power[16,17]. These results indicate a major role for genetic interactions in the heritability of complex traits. However, it is likely that the interactions uncovered to date represent only a fraction of the true genetic network.

Here, we show that both the power and interpretation of genetic interactions derived from association studies can be significantly improved through integration

with information about the physical architecture of the cell. We apply this integrative approach to an association study conducted in yeast, yielding a genetic network that complements, extends, and validates networks assembled through reverse genetic methods.

Chapter 1.4: Results

Chapter 1.4.1: Bi-clustering of marker pairs defines a network among genomic intervals

We analyzed a recent GWAS in yeast which analyzed a population of 112 segregants resulting from a cross of a laboratory S. cerevisiae strain with a wild isolate[5]. For each segregant, the states of 1,211 unique markers (genotypes) were mapped along with the expression profile of 5,727 genes (traits) (Table S1). To identify pairs of markers that genetically interact— i.e. for which the joint state of the marker pair was associated with one or more gene expression traits— we considered the method of Storey et al.[17] which provides the best marker pair for each expression trait, resulting in a set of 4,687 distinct marker-marker interactions (removing redundancies due to marker pairs that associate with multiple traits).

A preliminary examination of the genotype data showed few recombinations between neighboring markers, indicating that markers in close proximity were in linkage disequilibrium (LD). As a result, neighboring markers were often found to display similar patterns of interactions (Figure 1.1A). In much the same way that LD has allowed neighboring markers to be grouped into haplotype blocks[18], we reasoned that LD between neighboring markers could also be exploited to enhance marker-

marker interactions. To this end, we developed a bi-clustering algorithm to identify groups of marker-marker interactions that fall across common genomic intervals (Figure 1.1B; see Methods). We reasoned that bi-clustering the marker pairs might provide two distinct advantages: First, it allows many statistically insignificant marker-marker interactions to reinforce a single interval-interval interaction. Second, it leverages the structure between neighboring marker pairs to identify with greater precision the interval of DNA underlying the variance in a given trait.

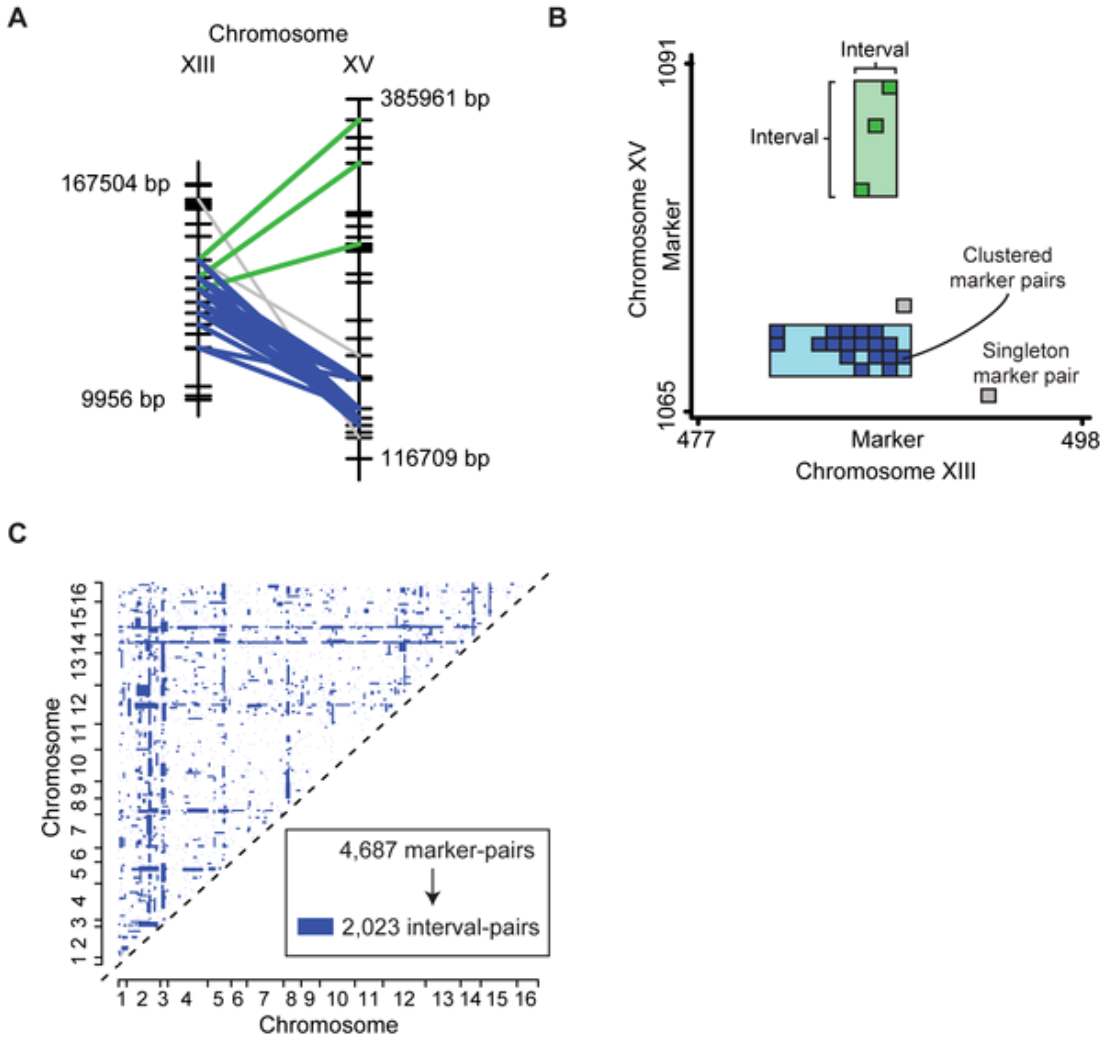


Figure 1.1: Using genome-wide association data to identify natural genetic interactions.
(A) Two interacting interval pairs (green and blue) which represent significantly dense groups of marker-marker interactions are shown. (B) A matrix view of the same genomic regions. The blue and green interval pairs appear as two rectangles. (C) The entire set of marker pairs was bi-clustered to form a set of high-confidence interval pairs (blue rectangles).

Applied to the marker pairs from Storey *et al.*, the bi-clustering procedure yielded a network of 2,023 interactions between 1,977 genomic intervals (Figure 1.1C). Of these, 695 interval pairs garnered support from multiple marker pairs (five on average). The remaining 1,328 interval pairs consisted of singleton marker-marker interactions, which were not found to cluster with any others. The complete network

of interval-interval interactions can be found in Table S2. We refer to this network as a natural genetic network since it is derived from natural rather than engineered mutations.

Chapter 1.4.2: Natural interactions define a map of functional links between protein complexes

A common interpretation of genetic interactions measured in reverse genetic screens has been the "between-complex" or "between-pathway" model, in which interactions are found to span pairs of protein complexes or functional annotations. Such complex-complex interactions have been instrumental in identifying synergistic or compensatory relationships[4,8,19]. Similarly, pairs of functional terms have served to identify functions that are cooperative or buffer one another[4].

To evaluate natural networks in this fashion, we examined all pairs of documented protein complexes (out of 302 in Gavin et al.[20] or the Munich Information Center for Protein Sequences [MIPS][21]) and all pairs of functional terms (out of 1,954 terms in the Gene Ontology [GO][22]) for enrichment for natural genetic interactions. As further described in Methods, we inspected all complex pairs and found 208 significant interactions in the natural network (False Discovery Rate<5%; Table 1). Similarly, we identified 17,714 significant interactions between functional terms. In contrast, far fewer results were found for complex or term interactions derived from the raw marker pairs of Storey et al. prior to bi-clustering these data into intervals

(Table 1). The full set of complex-complex and term-term interactions are available as a resource in Table S3 and on http://www.cellcircuits.org/qtlnet/.

Table 1.1: Correspondence of interval and marker pairs with complexes and functions.

			Between		Within	
	Nodes†	Edges‡	Complexes	Terms	Complexes	Terms
Storey *et al.*						
Bi-clustering*	1,977	2,023	208	17,714	0	12
Raw Marker Pairs	1,157	4,687	38	3,546	0	3
Full 2D ANOVA scan**						
Bi-clustering	1,387	964	0	19	0	0
Raw Marker Pairs	1,141	4,687	0	0	0	0
Synthetic Genetic Analysis	2,117	29,275	140	1,833	13	33

†Node definition: For Storey *et al.* and Full 2D ANOVA, nodes represent genomic intervals. For the synthetic network, nodes represent genes.
‡All cases report the number of distinct interactions in the network, removing redundancies due to marker pairs that associate with multiple traits (Storey *et al.*, Full 2D ANOVA) or gene pairs scoring positive in multiple data sets (Synthetic Genetic Analysis).
*These bi-clustered interval pairs were used to define the "Natural Network" explored in this work.
**We also considered an exhaustive scan of all marker pairs using two-way analysis of variance (ANOVA). The most significant 4,687 marker-marker interactions (Table S7) were taken to match the number of interactions from Storey *et al.* (Text S1). Both the raw marker-pairs and the bi-clustered interval network identified substantially fewer enrichments than the Storey *et al.* method.
doi:10.1371/journal.pgen.1000782.t001

Figure 1.2A shows a map of the 50 most significant complex-complex interactions. Because gene expression is the phenotypic trait, each complex-complex interaction is linked to a cluster of gene expression levels that it regulates (with each cluster containing an average of 287 genes). As the map integrates many traits simultaneously, it is distinct from previously-published genetic networks which have relied on cell viability as the single readout of interest. We found that two-thirds of the complex-complex interactions were linked to gene expression clusters that were highly functionally coherent (Figure 1.2A). In contrast, less than one one-hundredth of interval-pairs were found to influence a set of genes belonging to a single pathway or function. Thus, we conclude that integration of epistatic interactions with protein

complex maps helps to filter spurious interactions while simultaneously providing a putative mechanism for the pair-wise associations.

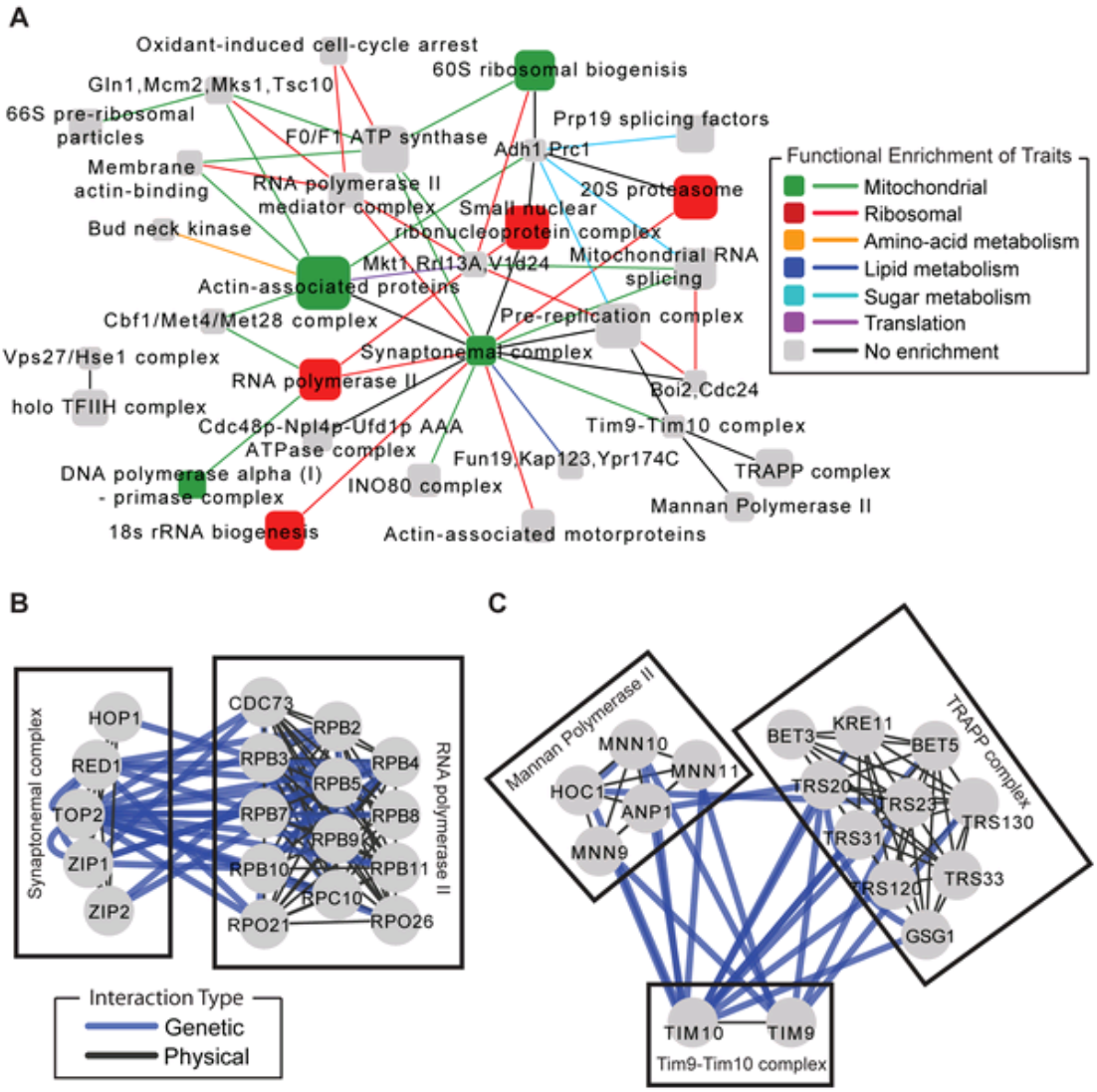


Figure 1.2: Natural genetic networks elucidate pathway architecture.
A global map of the top 50 complex–complex interactions found using the natural network. Each node represents a protein complex and each interaction represents a significant number of genetic interactions (False Discovery Rate<5%)[23]. We analyzed the set of gene expression traits associated with each complex-complex interaction for functional enrichment using the hypergeometric test. Nodes and edges are colored according to the functional enrichment of gene expression traits underlying the natural interactions (Bonferroni P′<0.05). Node sizes are proportional to the number of proteins in the complex. When available, nodes have been labeled with the common name of the complex. (B,C) Two specific examples of complexes spanned by dense bundles of natural genetic interactions.

As an illustrative example, Figure 1.2B shows the natural genetic interactions supporting a functional link between the synaptonemal complex and RNA Polymerase II. Mutations in the TOP2 gene of the synaptonemal complex have been shown to lead to higher levels of mitotic recombination in rDNA which can result in amplification and deletion of the rDNA array[24]. RNA polymerase II is responsible for the transcription of small nucleolar RNAs (snoRNAs) that physically and functionally interact with many other proteins required for ribosomal biogenesis[25]. Indeed, we found that the gene expression traits linked to this interaction were enriched for ribonucleoprotein complex biogenesis and ribosome biogenesis (both $P' = 10{-}8$ by hypergeometric test; P' is a Bonferroni corrected p-value).

Figure 1.2C centers on two of the interactions for the Tim9-Tim10 complex, an essential component of the TIM machinery responsible for the transport of carrier proteins from the cytoplasm to the inner mitochondrial membrane[26]. Tim9-Tim10 is genetically connected with Mannan Polymerase II and the TRAPP complex. Mannan Polymerase II is a component of the secretory pathway and is involved in lengthening the mannan backbone of cell wall and periplasmic proteins[27]; the TRAPP complex plays an important role in trafficking of proteins from the golgi to the cell periphery[28]. The abundant genetic interactions between Tim9-Tim10 and these two complexes suggest they may jointly influence the make-up of cell surface proteins, possibly through control of trafficking. Consistent with this hypothesis, disruption of mitochondrial function has been shown to influence cell wall composition, including levels of phosphopeptidomannans[29].

For comparison to the between-complex model, we also examined the natural genetic network for support for a "within-complex" model, in which single functional terms or complexes are enriched for genetic interactions among their member genes[4,8,19]. Searching across the 1,954 GO terms and 302 complexes, the natural network identified only 12 enriched GO terms and no significant complexes (Table 1 and Table S3). Thus, genetic interactions in naturally-derived networks are far less likely to occur within a single pathway than to span between pathways. This result mirrors what has been observed in analysis of reverse genetic interaction networks, particularly amongst interactions characterized as synthetic lethal or synthetic sick, which have been shown to interconnect different pathways that are functionally synergistic or redundant[19,30].

Chapter 1.4.3: Complementarity between natural and synthetic genetic networks

Next, we asked whether the natural genetic network had any direct overlap with "synthetic" networks derived using reverse genetic approaches such as SGA, dSLAM, or E-MAP platforms. To address this question, we considered four synthetic interaction networks: a network by Tong et al.[4] including comprehensive interaction screens for 132 genes using SGA, a genetic network governing DNA integrity identified using dSLAM[3], and E-MAPs centered on chromosomal biology[2] and RNA processing[1]. The combined network from these four sources consisted of 2,117 genes linked by 29,275 genetic interactions. As with the natural network, we confirmed that

interactions in the combined synthetic network were more likely to fall between functional terms and protein complexes than within them (Table 1 and Table S4).

To evaluate overlap, an interaction in the synthetic network was considered "supported" if the two genes mapped into two different intervals that were found to interact in the natural network. As shown in Figure 1.3A, the natural network supported on average 8.7% of interactions across the four synthetic networks as opposed to 5.7±0.5% expected by chance (Text S1). Thus, some regions are shared in common between natural and synthetic networks, although these regions appear to represent a minority of all genetic interactions.

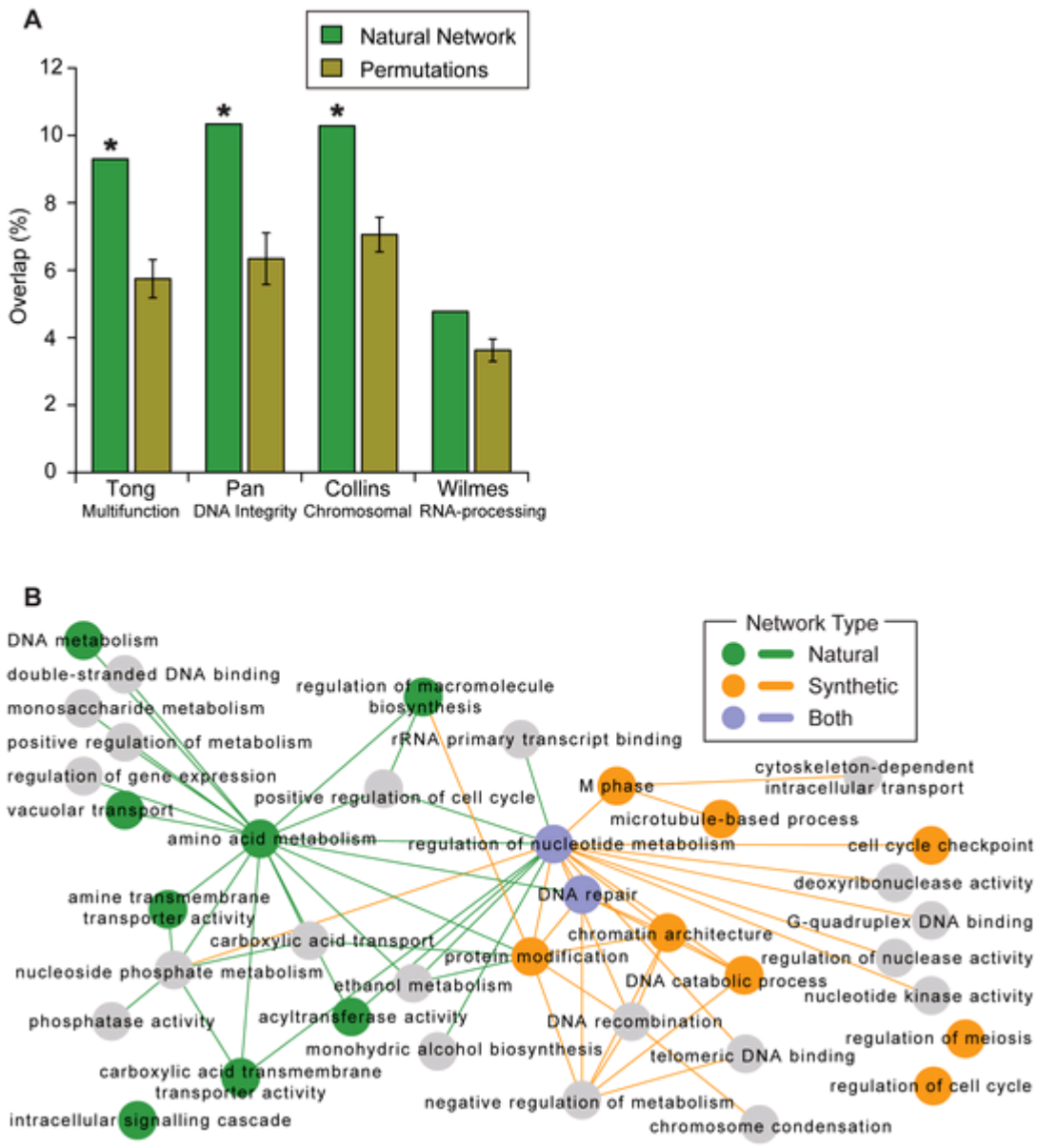


Figure 1.3: Comparison of the natural and synthetic networks.

(A) The overlap between the natural network and four previously-published synthetic genetic networks (Tong [4], Pan[3], Collins[2], Wilmes[1]) is shown as a percentage of the synthetic network size. An asterisk indicates significance at P<0.05. (B) A map of the functions and functional relationships supported by either the natural or synthetic networks. Each node represents a broad GO term, with colors (green, orange, blue) indicating terms that contain many within-term interactions (Text S1). Edges show the top 30 between-term interactions for each of the natural and synthetic networks. Two broad GO terms (regulation of nucleotide metabolism and DNA repair) contained many within-term interactions in both the natural and synthetic networks.

We found that these common genetic interactions took place among genes encoding basal transcriptional activators (“regulation of nucleotide metabolism”, Figure 1.3B) including components of RNA polymerase II, Kornberg's mediator complex, the holo TFIIH complex, INO80, SET3, and COMPASS (Figure 1.4A). The expression traits linked to these common interactions were for genes encoding the cytosolic ribosome ($P' < 10^{-47}$), cell cycle checkpoints ($P' < 10^{-15}$, including RAD9 and DDC1), and mitochondrial electron transport ($P' < 10^{-12}$). Thus, interactions that overlap between natural and synthetic genetic networks take place largely among core transcriptional activators and influence expression of core metabolic processes.

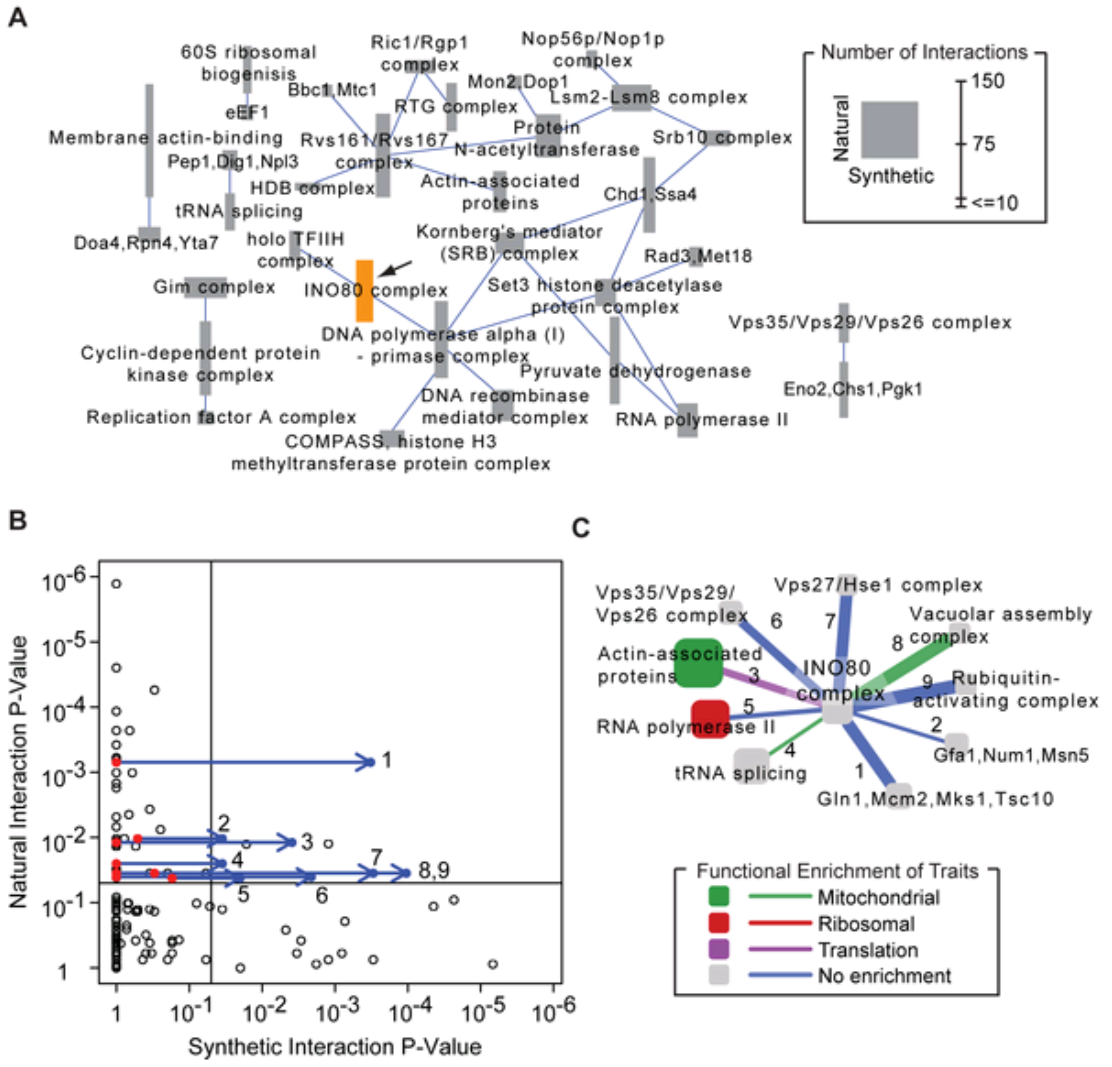


Figure 1.4: Guiding synthetic genetic screens using natural genetic networks.

(A) Complex-complex interactions common to both the natural and synthetic networks at a relaxed threshold of P<0.05. Many of these complexes, including INO80 (orange), have more coverage in the natural network (node height) than in the synthetic network (node width). (B) Each point in the scatter plot represents the significance of support for a possible complex-complex interaction with INO80 from the natural (y-axis) versus synthetic (x-axis) networks. Due to low coverage, comparatively few complex pairs have support in the synthetic network. New E-MAP data for INO80 support nine new complex-complex interactions predicted by the natural network (blue arrows). (C) A network of natural genetic interactions for INO80 validated by the new E-MAP. Functional enrichment for traits is shown as in Figure 1.2. The thickness of each link is proportional to its support in the new genetic interaction screen.

Chapter 1.4.4: Novel interactions of the INO80 complex as suggested by natural networks

One prominent complex highlighted by both natural and synthetic interactions was INO80, a multi-subunit ATP-dependent chromatin remodeling complex (Figure 1.4A). At its core is the Ino80 protein, an ATPase of the SNF2 family which functions as the catalytic subunit. Recent studies have demonstrated that INO80 chromatin remodeling activity contributes to a wide variety of pivotal processes, including transcription, DNA replication, and DNA repair[31–34]. Consistent with these processes, both the natural and synthetic networks supported interactions of INO80 with TFIIH and alpha(I)-primase. However, INO80 had far more interactions in the natural network than the synthetic one. This result is reflected in Figure 1.4A (large height versus width of the INO80 node) and more explicitly in Figure 1.4B, which plots the p-values in the natural versus synthetic network for all complex pairs involving INO80. This plot suggests that the reason for few synthetic interactions is lack of coverage: most complex pairs (82%) have simply not yet been tested for interaction using reverse genetic screens, placing them at a significance score of P = 1 (i.e., on the y-axis of Figure 1.4B).

To fill this gap, we genetically analyzed three genes encoding members of the INO80 complex (Arp8, Ies3, Nhp10) using the quantitative E-MAP approach. Complete genomic deletions of each gene were screened against a standard array of 1,536 mutants to select double mutant combinations whose growth rates were slower or faster than expected (Methods). This screen uncovered 496 novel genetic

interactions (Table S5) supporting 20 complex-complex relationships ($P<0.05$; Table S6). Nine of the complex-complex interactions were also supported by the natural network, including interactions with four complexes (tRNA splicing, RNA polymerase II, Actin-associated proteins, and the Vps35/Vps29/Vps26 complex) that were already present in the common complex interaction map (see Figure 1.4B and 1.4C).

The relationships identified here implicate a number of novel links between INO80-mediated chromatin remodeling and a wide range of important cellular processes. For example, numerous genetic interactions were identified between INO80 and RNA Polymerase II. There is substantial evidence demonstrating that the rate of transcriptional elongation by RNA Polymerase II is reduced in the presence of nucleosomes and requires chromatin-modifying activities[35]. Since INO80 has been shown to mobilize/remove nucleosomes[33,36], this functional link may indicate that the two complexes co-operate: INO80 may exchange histones at a particular location to facilitate transcriptional elongation by RNA polymerase II. Indeed, while this manuscript was in review, a new report has implicated a role for INO80 in histone redeposition during RNA polymerase II-mediated transcription of stress-induced genes[37].

Four of the nine novel INO80 interactions are involved in various aspects of vacuolar protein degradation including transport of hydrolases to the vacuole (Vps35/Vps29/Vps26 complex and Vps27/Hse1 complex), vacuole biogenesis (Vacuolar assembly complex), and targeting of proteins for degradation (Rubiquitin-activating complex). Given INO80's role in transcription[33], the new interactions

suggest that these complexes work in tandem to regulate the expression level of certain proteins, with INO80 controlling the level of transcription and these four complexes controlling the rate of protein degradation. This work serves as an example of how the broad coverage in the natural network can be used to focus future genetic screens and provide the basis for many mechanistic follow-up studies.

Chapter 1.5: Discussion

Currently, mapping genetic interactions using GWAS faces two major challenges: a lack of statistical power for finding genotype-phenotype associations, and a lack of tools for understanding the molecular mechanisms behind the associations found to be significant[14,15,38]. In this study, we have demonstrated that such challenges can be partly overcome by (1) accounting for bi-cluster structure in the data and (2) by integrating genetic interactions derived from GWAS with protein complexes and functional annotations. The result is a map of protein complexes and pathways interconnected by dense bundles of genetic interactions, which raises statistical power and provides biological context to the genetic interactions uncovered in natural populations.

Despite exhibiting some overlap (8.7%), there was also much divergence between the natural and synthetic networks. Such divergence might be explained by a number of factors. First, the two types of genetic networks have major differences with respect to coverage and power. Natural networks are based on genome-wide variations and thus nearly all gene pairs are tested for pairwise interaction— i.e., the

coverage of gene pairs is practically complete. This large coverage comes at the price of low statistical power: gene association studies are limited by the number of individuals that can be surveyed which, in turn, limits the power of natural genetics to detect any given genetic interaction. On the other hand, a reverse genetic interaction screen explicitly tests the growth rate of gene pairs, with high power to detect interaction. However, the set of gene pairs that can be tested in a single study is limited by the throughput of the screening technology. The synthetic genetic network used here was a combination of four such studies which collectively cover approximately 5% of yeast gene pairs. Future efforts may seek to complement the coverage of reverse genetic screens by using natural genetics, or to improve the power of gene association studies through focused reverse genetic analysis. Here, we have demonstrated this concept by expanding the coverage of the synthetic network around the INO80 complex, based on the conserved interactions we found for this complex in both types of networks.

Even with equivalent coverage and power, the two types of network would still likely diverge due to their different means of perturbation. The natural network is driven by variations in genome sequence including SNPs, repeat expansions, copy number variations, and chromosomal rearrangements which lead to a variety of effects on gene function such as hypo- and hypermorphic alleles, null alleles, and so on. In contrast, synthetic networks predominantly consist of complete gene deletion events, which are rarely experienced in nature and lead exclusively to null alleles.

A final difference is phenotype— the natural and synthetic networks in this study differ markedly in the underlying phenotypic traits they have measured, relating to gene expression versus cell growth, respectively. It is important to note, however, that the differences in traits are specific to the currently available data sets. They are not inherent to either mapping approach, and in general one can imagine synthetic genetic interactions related to gene expression (see Jonikas et al. for a recent example[39]) or natural interactions related to a single phenotypic trait such as cell viability or disease (which in fact describes the majority of GWAS data generated to-date for humans)[7].

Despite all of these differences, we did observe a significant number of natural and synthetic genetic interactions in common. It is tempting to speculate that these common interactions might share certain characteristics with regard to cellular function. In particular, we found that natural interactions also present in the synthetic network were linked to expression levels of ribosomal genes as well as to core components of respiration and cell cycle. Several studies have noted a correlation between the expression levels of ribosomal or mitochondrial genes and growth rate[40,41]. Thus, the overlap between natural and synthetic interactions seems to occur among genes that strongly influence expression traits related to growth.

A common issue in association studies, known as the "fine mapping problem"[42,43], is that a strongly associated marker will fall near many candidate genes, leaving it ambiguous as to which of these candidates is the causal factor. Numerous methods have been developed to refine or prioritize these candidates, often through

incorporation of orthogonal information[44]. An extension of this problem applies to marker-marker interactions, which typically implicate one of many possible pairs of genes. Here, we have mitigated this problem by summarizing markers into protein complexes and functional terms. However, ambiguities can still arise in cases where several complex-complex interactions are supported by the same underlying set of marker pairs. Since it is likely that only one of these interactions is causally linked to phenotype, further work may be necessary to prioritize these candidates. It is important to note, however, that fine-mapping issues will be less of a concern in humans than in yeast, given the higher density of available markers which will improve the resolution in identifying causal genes.

In summary, we have demonstrated that the logical framework developed for analysis of synthetic genetic networks can also be readily applied to natural genetic networks. Biologically and clinically, the clear and immediate application is towards the analysis of genome-wide association studies in humans. Many diseases, both common and rare, have so far been opaque to genome-wide association analysis[45]. The key question will be whether, using integrative maps such as those developed here, they can become less so.

Chapter 1.6: Methods

Chapter 1.6.1: Marker pair bi-clustering

An interval is defined as a set of one or more contiguous markers along the chromosome. A pair of intervals induces a set of m tested marker pairs of which k

pairs are found to interact, drawn from a total genome-wide pool of N tested marker pairs of which n are found to interact. An exhaustive genome-wide scan is performed to identify interacting interval pairs, i.e. those that are enriched for marker-marker interactions, as follows. The counts (m, k) are tallied for all possible pairs of intervals (up to a maximum of 60 markers per interval) using a recursive algorithm in which the entire space of marker pairs is represented as an upper-triangular matrix A with each row and column denoting a marker. An interval pair is represented by a submatrix Ai,j,a,b, where i,j are the starting row and column indices and a,b are the dimensions of the submatrix. The number ki,j,a,b of interacting marker pairs in a submatrix is determined using the formula:

An identical formula is used to count the number of tested marker pairs in each interval pair (substitute m for k). Following computation of the (m, k) counts, every interval pair is assigned a p-value of enrichment for marker-marker interactions based on the four parameters m, k, N, n using the hypergeometric distribution. The natural network is then assembled in an iterative fashion, where the most significant interval pair is selected from among all possible interval pairs, after which all interval pairs which contain any overlapping marker pairs (interacting or non-interacting) are removed from consideration. The process is repeated until there are no interval pairs remaining, which ensures that the final set of interval-interval interactions comprising the natural network is disjoint.

Chapter 1.6.2: Comparison of bi-clustering to a naïve algorithm

We considered that the improved performance of bi-clustering might be non-specific, i.e., that simpler methods for expanding marker-marker pairs to form genomic intervals might perform equally well. As one possibility, we compared the bi-clustering approach to a naïve algorithm for generating interval-interval interactions, in which raw marker pairs were expanded to encompass the nearest x neighboring markers on either side. However, as shown in Figure S1 this naïve expansion method performed substantially worse than bi-clustering at identifying term-term or complex-complex interactions, for any choice of x, suggesting that bi-clustering identifies more appropriate interval boundaries for each natural genetic interaction.

Chapter 1.6.3: Mapping genes to intervals

The chromosomal coordinates of open reading frames (ORFs) for all yeast genes were obtained from the Saccharomyces Genome Database[46]. Each gene was assigned to all markers found within its ORF and to the nearest marker within a window of x = 100 kb on either side (Figure S2). This mapping procedure resulted in a discrete number of genes mapped to a given marker. Intervals were mapped to all genes assigned to their constituent markers, again resulting in a discrete number of genes mapped to an interval.

The complex-complex interactions identified in the natural network were robust to the particular choice of window size x. We varied x over a range of distance thresholds from 0 to 100 kb. As shown in Figure S3, the resulting complex-complex

interactions implicated by the natural network had a high degree of overlap with the results obtained using the original mapping procedure.

Chapter 1.6.4: Enrichments of interactions within and between complexes and terms

A within-complex (within-term) model is defined as the set of all gene pairs falling within a given physical complex (functional GO term). A between-complex (between-term) model is defined as the set of all gene pairs that span two complexes (terms), such that one gene belongs to the first complex, the other gene belongs to the second complex, and neither gene belongs to both. For each model we compute k, the number of gene pairs "supported" (see main text) by the network. The significance of this support is assessed using the hypergeometric distribution, governed by k and three additional parameters:

n. The total number of gene pairs induced by the model.

m. The total number of gene pairs having support in the entire network.

N. The total number of gene pairs in the tested space of the entire network.

Counts for all four parameters are based only on pairs of genes found in the corresponding space of interactions tested by the network and covered by the given annotation set (complexes or terms). Further details are given in Text S1. All models are visualized using Cytoscape[47].

Chapter 1.6.5: Removing the effects of non-random gene order on annotation enrichment

The above enrichment tests assume independence of genetic interactions from protein complexes and functional terms. However, intervals in the natural network typically cover several consecutive genes, which are more likely to be of similar function than genes chosen at random[48]. To correct for this effect, each complex/term annotation is assigned a score Pmin∈[0, 1] measuring the degree to which its member genes are clustered [Pmin → 0] versus dispersed [Pmin → 1] along the genome (see Text S1 for more details). Annotations with Pmin<pT are removed from further consideration. We use a stringent threshold of pT = 0.1 for physical complexes and pT = 0.3 for functional terms resulting in less than one erroneous complex-complex or term-term interaction identified in randomized networks (Figure S4 and Figure S5). Further details regarding the randomization procedure is provided in Text S1. A list of the complexes used in this study is provided in Table S8.

Chapter 1.6.6: INO80 Epistatic Mini-Array Profile (E-MAP)

The arp8Δ, nhp10Δ, and ies3Δ knockout strains were constructed and E-MAP experiments were performed as described previously[49]. The array used to generate the double-knockout strains contained 1,536 strains involved in chromatin metabolism (including chromatin remodeling, repair, replication, and transcription) as well as global cellular processes like protein trafficking and mitochondrial metabolism (see Table S5). Genetic interaction scores were computed as described previously[9].

Chapter 1.7: Acknowledgments

Hannum G, Srivas R, Guénolé A, van Attikum H, Krogan NJ, Karp R, Ideker T. (2009) Genome-Wide Association Data Reveal a Global Map of Genetic Interactions among Protein Complexes. PLoS Genet 5(12): e1000782. doi:10.1371/journal.pgen.1000782

Gregory and Rohith contributed equally to this work. Conceived and designed the experiments: GH RS RMK TI. Performed the experiments: GH RS TI. Analyzed the data: GH RS RMK TI. Contributed reagents/materials/analysis tools: GH RS AG HvA NJK TI. Wrote the paper: GH RS TI.

We thank Sourav Bandyopadhyay for numerous comments and suggestions. Tune H. Pers, Karen Kapur, and Ryan Kelley provided helpful reviews of the manuscript.

CHAPTER 2: ASSEMBLING GLOBAL MAPS OF CELLULAR FUNCTION THROUGH INTEGRATIVE ANALYSIS OF PHYSICAL AND GENETIC NETWORKS

Chapter 2.1: Abstract

To take full advantage of high-throughput genetic and physical interaction mapping projects, the raw interactions must first be assembled into models of cell structure and function. PanGIA (for physical and genetic interaction alignment) is a plug-in for the bioinformatics platform Cytoscape, designed to integrate physical and genetic interactions into hierarchical module maps. PanGIA identifies 'modules' as sets of proteins whose physical and genetic interaction data matches that of known protein complexes. Higher-order functional cooperativity and redundancy is identified by enrichment for genetic interactions across modules. This protocol begins with importing interaction networks into Cytoscape, followed by filtering and basic network visualization. Next, PanGIA is used to infer a set of modules and their functional inter-relationships. This module map is visualized in a number of intuitive ways, and modules are tested for functional enrichment and overlap with known complexes. The full protocol can be completed between 10 and 30 min, depending on the size of the data set being analyzed.

Chapter 2.2: Introduction

Genetic interactions are defined as functional relationships between genes that result when the phenotypic effect of one gene is altered by one or several other genes[8,50]. Such interactions have been used to uncover pathway architecture in model

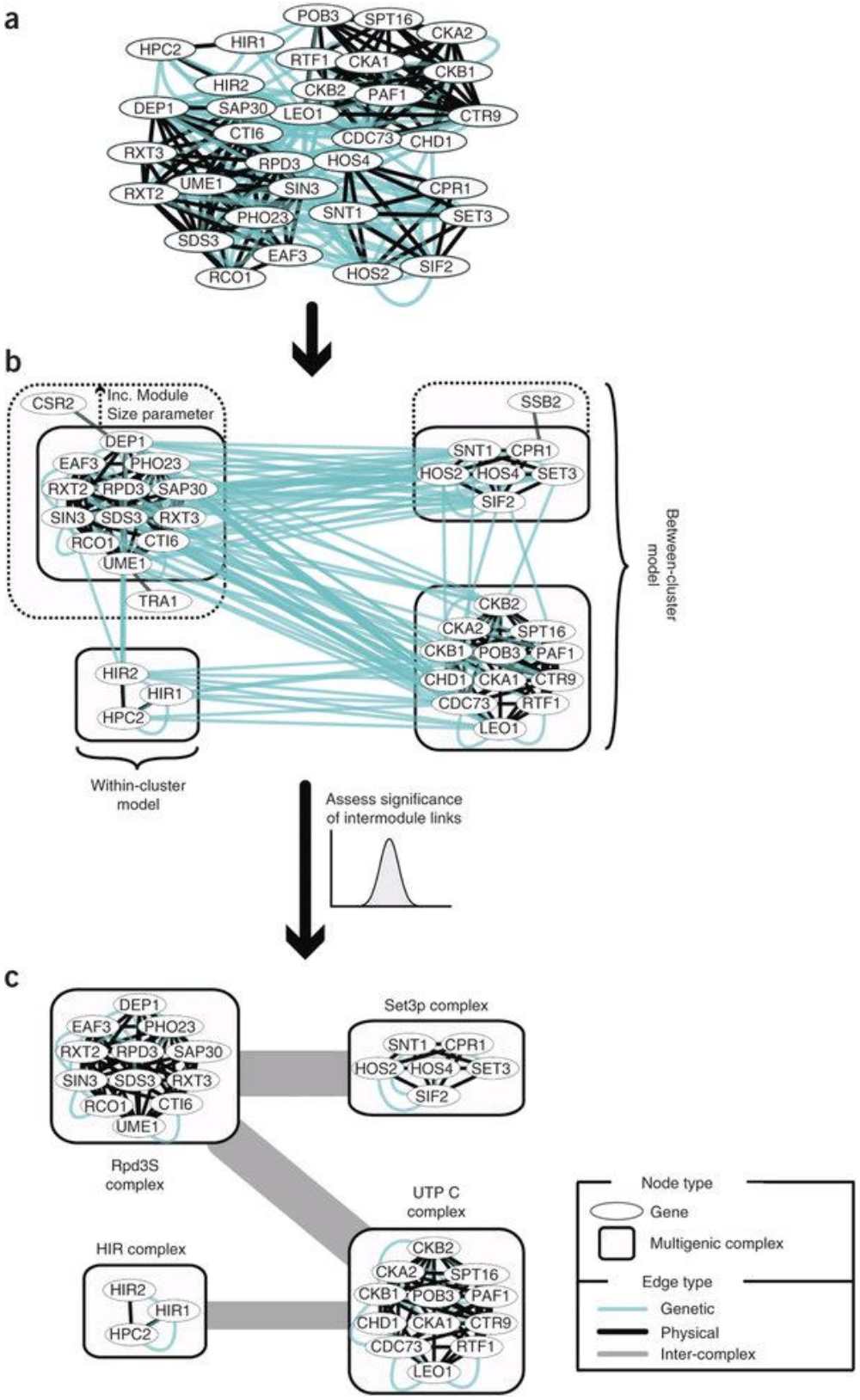
organisms[2,4,11,51]. In humans, genetic interactions are thought to influence numerous phenotypes of interest, from expression[6] to complex diseases[7] to drug resistance[52]. Recently, a number of technologies such as synthetic genetic arrays[4,9,49,53] and heterozygote diploid-based synthetic lethality analysis with microarray[3] have facilitated the rapid screening of genetic interactions in model organisms. In human cell lines, combinatorial RNA interference screening technologies have begun to show promise in uncovering genetic interactions[54,55]. As a result of these high-throughput technologies, the amount of genetic interaction data available in the public domain has increased rapidly. As of December 2010, the BioGRID interaction database housed nearly 175,000 genetic interactions spanning 11 different species[56].

Interpreting the functional significance of each genetic interaction remains a daunting task. One promising solution has been to interpret genetic interactions in the context of their relationships to physical protein-protein interactions (Figure 2.1a)[19,30,57,58]. At least two distinct models have been put forth to reconcile genetic and physical interactions. The 'within-cluster' model seeks to identify clusters of proteins that are enriched for both physical and genetic interactions (Figure 2.1b). We refer to such clusters of proteins and the interactions occurring among them as a module. Modules are often interpreted as functional protein complexes[4,19,30,57] or signaling pathways[52]. In contrast, the 'between-cluster' model seeks genetic interactions that are enriched across two clusters of interacting proteins (Figure 2.1b). Such intermodule links have been shown to identify synergistic or compensatory relationships between protein complexes or signaling pathways[2,19,58]. Figure 2.1c shows an example module

map consisting of four modules connected by three intermodule links. The genes in each of these four modules are associated with a strong within-cluster signal, and, furthermore, they coincide with known Saccharomyces cerevisiae physical complexes (Figure 2.1c). Set3p and Rpd3s are both histone deactylase complexes involved in transcriptional regulation. The Hir complex functions in replication-independent nucleosome assembly, whereas the UTP-C complex is a component of the 90S preribosome. The intermodule link between Set3p and Rpd3s suggests a functional synergy between the two complexes. Consistent with this hypothesis, several studies have illustrated that the two are jointly responsible for the the activation of DNA damage response genes via the recruitment of RNA Polymerase II (ref. 21).

Figure 2.1: Overview of PanGIA's method for identifying a module map of cellular function from physical and genetic networks.

(**a**) PanGIA takes as input a physical and genetic network. Black edges refer to physical interactions, whereas turquoise edges refer to genetic interactions. (**b**) Both within-cluster and between-cluster models are identified using the physical and genetic network. A within-cluster model or module consists of a set of genes connected by a large number of physical and genetic interactions. In this example four within-cluster models are identified. A between-cluster model or intermodule link consists of two within-cluster models spanned by a bundle of genetic interactions. Here, five putative between-cluster models have been identified. The size of within-cluster models can be controlled via the Module Size parameter. Higher values of the Module Size parameter lead to larger complexes (denoted by the dashed line). (**c**) If quantitative interaction data have been made available, the significance of each between-cluster model can be assessed. Only significant intermodule links are displayed in the final module map (three of the five putative intermodule links are significant in this example). The thickness of the line reflects the score of the intermodule link, which is based on the number of physical and genetic edges spanning the two modules. If a biological annotation set is provided, PanGIA will check the overlap between the set of genes comprising the annotation and the set of genes comprising each module. If the overlap exceeds a user-specified threshold, the module will be labeled with the name of the annotations. Here, all four modules overlap with known complexes and are labeled accordingly.


a
POB3
SPT16
CKA2
HPC2
HIR1
RTF1
CKA1
CKB1
HIR2
CKB2
PAF1
DEP1
SAP30
LEO1
CTR9
CTI6
CDC73
CHD1
RXT3
RPD3
HOS4
UME1
SIN3
CPR1
RXT2
PHO23
SNT1
SET3
SDS3
EAF3
HOS2
SIF2
RCO1
b
CSR2
Inc. Module
Size parameter
SSB2
DEP1
SNT1
CPR1
EAF3
PHO23
HOS2
HOS4
SET3
RXT2
RPD3
SAP30
SIF2
SIN3
SDS3
RXT3
RCO1
CTI6
UME1
TRA1
CKB2
CKA2
SPT16
CKB1
POB3
PAF1
CHD1
CKA1
CTR9
CDC73
RTF1
LEO1
HIR2
HIR1
HPC2
Between-cluster
model
Within-cluster
model
Assess significance
of intermodule links
c
DEP1
Set3p complex
EAF3
PHO23
SNT1
CPR1
RXT2
RPD3
SAP30
HOS2
HOS4
SET3
SIN3
SDS3
RXT3
SIF2
RCO1
CTI6
UME1
Rpd3S
complex
UTP C
complex
HIR complex
CKB2
HIR2
CKA2
SPT16
HIR1
CKB1
POB3
PAF1
HPC2
CHD1
CKA1
CTR9
CDC73
RTF1
LEO1
Node type
Gene
Multigenic complex
Edge type
Genetic
Physical
Inter-complex

Several methods have been previously published[19,30,57,59] for analyzing interactions to identify both within-cluster and between-cluster functional organization. However, these methods have not yet been made available through a publicly accessible software package. Here we introduce a novel software tool, PanGIA, along with a general bioinformatics protocol for integrative analysis of genetic interactions. PanGIA implements a previously published framework[58] as a plug-in for the open-source network analysis platform, Cytoscape[47,60], and allows the user to easily generate maps of modules and module inter-relationships from genetic and physical interaction data (see Figure 2.1 for an overview). A number of options are available to the user for constructing and visualizing the resulting module map. PanGIA is built on the new Cytoscape 2.8 architecture[61], which features the ability to view and manipulate nested networks, thereby enabling the user to explore both the global map as well as individual modules in an intuitive manner. Finally, individual modules can be interrogated using a number of functional enrichment options.

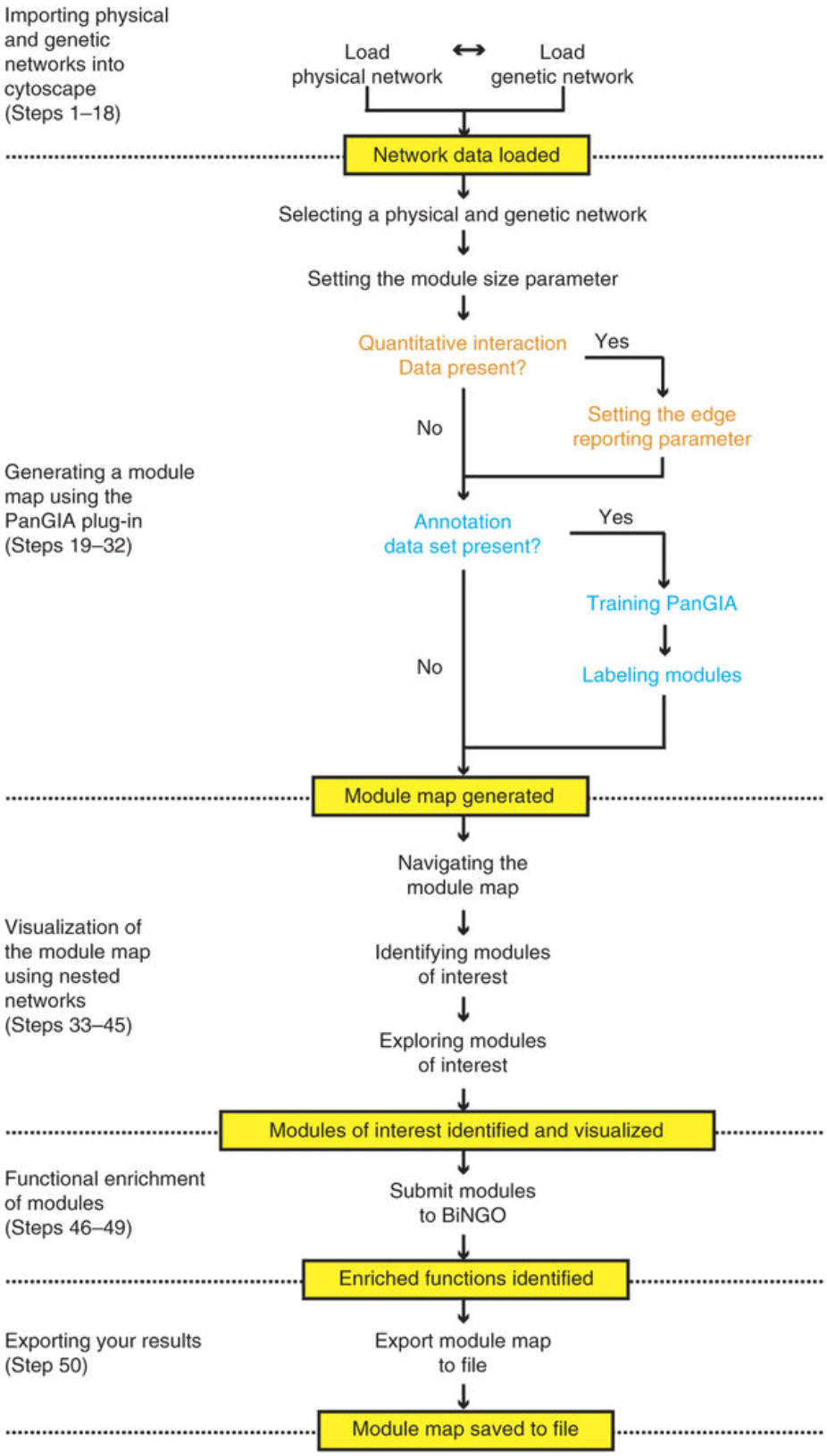
The computational workflow presented here has been used in the analysis of genetic networks centered on genes involved in chromosomal biology[2,58], RNA processing[1], secretory pathways[51] and DNA damage response[52]. This analysis has also been used in comparing genetic networks across two different species[10]. In each case, the module maps generated have helped to identify novel pathways as well as new components and functions for existing complexes[10,19,30,52,58]. While this workflow has proven useful in the analysis of numerous genetic interaction data sets, the module search process works best when there is a high density of protein and genetic

interactions among the set of genes being studied. For species in which there is a scarcity of either genetic interaction or physical interaction data, this protocol may not identify a significant number of modules or intermodule relationships. This limitation will become less relevant as large-scale interaction screens continue to populate the scientific databases.

This protocol is divided into five basic sections (Figure 2.2). The first section, 'Importing physical and genetic networks into Cytoscape', describes the available sources of interaction data and means of acquiring these data within Cytoscape. Second, 'Generating a module map using the PanGIA plug-in' covers the use of the PanGIA plug-in and is further divided into four subsections covering the various aspects of its use ('Selecting a physical and genetic network', 'Setting the module size and edge reporting parameters', 'Training PanGIA' and, finally, 'Labeling modules'). The third section, 'Visualization of the module map using nested networks', introduces ways in which the user can navigate and visualize the resulting module map. Fourth, 'Functional enrichment of the modules' illustrates methods to identify enriched biological functions and pathways among the identified modules. Finally, 'Exporting the results' covers the various ways in which the module map can be exported from Cytoscape for further analysis or for inclusion as figures in a publication.

Figure 2.2: Outline of the protocol.

Analyses listed in black indicate required steps in the protocol. Analyses listed in orange represent optional steps, which may be performed if quantitative interaction data are present; those listed in light blue are optional steps, which may be performed if a biological annotation data set is present. The yellow boxes indicate the desired outcome at the end of each major section in the protocol.


Importing physical and genetic networks into cytoscape (Steps 1–18)
Load physical network
Load genetic network
Network data loaded
Selecting a physical and genetic network
Setting the module size parameter
Quantitative interaction Data present?
Yes
Setting the edge reporting parameter
No
Generating a module map using the PanGIA plug-in (Steps 19–32)
Annotation data set present?
Yes
Training PanGIA
Labeling modules
No
Module map generated
Navigating the module map
Visualization of the module map using nested networks (Steps 33–45)
Identifying modules of interest
Exploring modules of interest
Modules of interest identified and visualized
Functional enrichment of modules (Steps 46–49)
Submit modules to BiNGO
Enriched functions identified
Exporting your results (Step 50)
Export module map to file
Module map saved to file

Importing physical and genetic networks into Cytoscape:

This section of the PROCEDURE (Steps 1–18) describes the various ways in which a physical or genetic network can be imported for analysis into Cytoscape. A previous protocol has outlined the various file formats Cytoscape can recognize as well as provided detailed instructions on how each file type can be imported[60]. The present protocol will instead focus on importing networks in a tab-delimited format (Box 1). Table 1 provides examples of several different databases from which interaction data (both genetic and physical) can be downloaded in a tab-delimited format for over 50 organisms.

Table 2.1: List of databases of physical and genetic interaction data.

Database Name	URL	# of Organisms Covered	Physical Interaction Data Available?	Genetic Interaction Data Available?	Quantitative Interaction Data Available?
STRING	string-db.org	630	yes	no	yes
DIP	dip.doe-mbi.ucla.edu/dip/Main.cgi	372	yes	no	yes
IntAct	www.ebi.ac.uk/intact/main.xhtml	305	yes	no	yes
Consensus PathDB	cpdb.molgen.mpg.de	3	yes	no	no
BioGRID	thebiogrid.org	18	yes	yes	no
MINT	mint.bio.uniroma2.it/mint/Welcome.do	30	yes	no	yes

Generating a module map using the PanGIA plug-in:

Selecting a physical and genetic network. This section of the PROCEDURE (Steps 19–23) describes the steps necessary to select which physical and genetic networks are to be analyzed. At this point, PanGIA is fully configured and the module search process can be initiated. However, PanGIA is designed with four optional features designed to fine-tune and enhance the search process. We describe these optional features in the subsequent sections.

Setting the module size and edge reporting parameters (Steps 24–26):

The first optional feature is the 'module size' parameter. This parameter helps to control both the size and number of modules by rewarding the formation of larger modules. Thus, higher values of this parameter results in the formation of larger, but fewer modules. Lower values produce the opposite effect (Figure 2.1b). It is recommended that the module size parameter initially be left at the default value. If

the resulting module map contains very large modules, the module size parameter can be suitably altered and the module search process re-run to produce smaller and more biologically meaningful modules.

The second optional feature is dependent on the presence of quantitative genetic interaction data. Many of the recent experimental technologies for measuring genetic interactions go beyond reporting interactions in a simple binary format (interacting or noninteracting) and provide some measure of confidence in a given interaction. For example, in the synthetic genetic array technology[53] and a recent variant called epistatic mini-array profiles[2,9], each double mutant is assigned a quantitative signed score, where positive scores indicate that the double mutant grew better than expected (e.g., suppression) and negative scores indicate pairs for which the double mutant grew worse than expected (e.g., synthetic sick or synthetic lethal)[9,53]. Table 1 outlines numerous databases that contain quantitative interaction data.

If quantitative genetic interaction data are provided, each intermodule link can be assessed for significance. A P value is assigned by comparing the sum of the interaction confidence values for all genetic interactions spanning two modules (i.e., intermodule link) to a distribution of the sums of confidence values of an equal number of genetic interactions drawn at random[58] (Figure 2.1c). The edge reporting parameter serves as a threshold; only those interactions with a P value less than this threshold are displayed in the final module map. By default, this parameter is set to 0.1, thus displaying only those intermodule links with $P < 0.1$.

Training PANGIA (Steps 27–29):

The next optional feature relies on the presence of a biological annotation set. Examples of an annotation set that can be used include physical complexes, signaling pathways, metabolic pathways or even broad biological processes. Table 2 provides a list of databases where an annotation set can be downloaded for a range of different organisms.

Table 2.2: Examples of databases from which to obtain annotation data.

Database Name	URL	# of Organisms Covered	Annotation Type
Gene Ontology (GO)	www.geneontology.org/GO.downloads.annotations.shtml	48	Physical complexes, biological processes, signaling pathways, metabolic pathways
MIPS CORUM	mips.helmholtz-muenchen.de/genre/proj/corum	3	Physical complexes
KEGG	www.genome.jp/kegg/pathway.html	833	Metabolic pathways, signaling pathways
CYC2008	wodaklab.org/cyc2008/	1 (*S. cerevisiae*)	Physical complexes
SGD Pathways	pathway.yeastgenome.org	1 (*S. cerevisiae*)	Metabolic pathways
MetaCyc	metacyc.org	2000	Metabolic pathways
Reactome	www.reactome.org	20	Metabolic pathways

The optional training procedure built into PanGIA is designed to help identify modules that are more likely to be biologically relevant, i.e., modules that contain

genes that operate in the same complex or biological process. By default, the module search process is designed to identify sets of genes that are densely connected by physical and genetic interactions. However, some interactions can be given more or less influence based on their quantitative score. PanGIA can determine how likely a certain interaction (either physical or genetic) is to connect two genes within a known complex or biological process using an existing annotation set. Examples of such a set include physical complexes (e.g., INO80 complex), signaling pathways (e.g., the mitogen-activated protein kinase (MAPK) pathway), metabolic pathways (e.g., glycolysis) or biological processes (e.g., DNA damage response genes). Using this annotation set, PanGIA assigns each interaction a weight based on the unsigned logistic regression of all interaction confidence scores of a given type (physical, genetic) against its proteins' co-membership in an annotation. If no quantitative scores are available, PanGIA uses logistic regression to assign a constant confidence score for all interactions of a given type. For specific details regarding the regression procedure, please see Bandyophadyay et al.[58]. The module search process will now seek to identify sets of genes that are connected by highly weighted physical and genetic interactions. As the weight of an interaction corresponds to how likely it is to connect two genes belonging to the same physical complex or pathways, the modules identified will contain genes that are functionally similar.

Labeling modules:

The genes composing a module may function in the same biological process or encode members of the same protein complex. If a biological annotation set is

provided, PanGIA will check to see if the module gene set overlaps with the annotation gene set. Here overlap is defined using the Jaccard similarity coefficient (intersection/union), which ranges from 0 (no overlap) to 1 (perfect overlap). If the Jaccard coefficient exceeds a user-specified threshold, then the module will be labeled with the name of the annotation in the final module map (Figure 2.1c). This PROCEDURE subsection (Steps 30–32) covers how this labeling feature can be enabled and provides instructions on how to set the overlap threshold.

Visualization of the module map using nested networks:

PanGIA is built on the new Cytoscape 2.8 architecture, which features the ability to view nested networks (i.e., each node in a network can represent an entire subnetwork). Instructions are provided for laying out the network of modules and intermodule links and for probing individual modules. This PROCEDURE section is divided into three subsections, 'Navigating the module map' (Steps 33–35), 'Finding modules of interest' (Step 36) and 'Exploring modules of interest' (Steps 37–45), which cover the various ways in which both the module map and individual modules can be interrogated.

Functional enrichment of the modules:

Modules will often contain genes of unknown function. One way to dissect the function of modules uncovered in this workflow is to examine if they are substantially enriched for any functional annotations. This can be used to identify new components of existing complexes or to identify entirely new physical complexes or pathways[2,19,58]. This PROCEDURE section (Steps 46–49) outlines the steps for

checking for enriched Gene Ontology (GO) functional terms[22] using the BiNGO plug-in[62].

Exporting your results:

This PROCEDURE section (Step 50) covers the various options for exporting the resulting module map.

Chapter 2.3: Materials

Equipment:

Personal computer with Internet access and an Internet browser.

Equipment setup:

Hardware requirements:

PanGIA hardware requirements depend on the size of the physical and genetic networks to be imported and analyzed. For networks containing up to 200,000 edges, we recommend a 2.0-GHz CPU or higher, a medium-end graphics card, 150 MB of available hard disk space and at least 2 GB of free physical RAM. If you are analyzing very large networks (>500,000 interactions), at least 8 GB of free physical RAM is recommended. To view the modular map produced by PanGIA, we recommend a monitor with a minimum screen resolution of 1024 × 768.

Operating system:

PanGIA and Cytoscape are supported on Windows (XP, Vista and Windows 7), Mac OS X (version 10.6 (Snow Leopard) or higher) and Linux.

Java standard edition:

Version 1.6 or higher is required (can be downloaded from http://www.java.com/).

A three-button mouse:

This is recommended (but not required) as an aid in navigating the module map.

Cytoscape v2.8.0:

PanGIA requires Cytoscape version 2.8.0 or higher. The steps for downloading and installing the latest version of Cytoscape can be found in a previously published protocol[24] or online at http://www.cytoscape.org/documentation_users.html.

Plug-ins:

The analysis capabilities of Cytoscape are expandable and extensible through add-on software packages called plug-ins. This protocol requires the installation of four plug-ins: PanGIA, BiNGO[62], Enhanced Search[63] and CyThesaurus[64]. Instructions for installing these plug-ins are outlined in PROCEDURE Steps 2–4.

MeV version 4.6 or higher:

MeV or MultiExperiment Viewer[65] is an integrated toolkit for clustering and visualizing large-scale genomic data. This protocol uses MeV to view modules as a hierarchically clustered heat map. Instructions for downloading and installing MeV can be found at http://www.tm4.org/mev/.

Data files:

PanGIA requires both a physical and genetic network in a tab-delimited format (Box 1). Sample protein and genetic interaction networks are provided as examples to illustrate the protocol. The physical interaction network (Supplementary Data 1) was taken from a recent integration of two high-throughput protein interaction screens[66]. Each physical interaction was assigned a Purification Enrichment score, with larger values representing greater confidence in the physical interaction. The genetic interaction network (Supplementary Data 2) was obtained from a large epistatic mini-array profile screen, which measured all possible genetic interactions among 743 genes involved in yeast chromosomal biology[2]. Each genetic interaction was assigned an S-score representing both the magnitude and confidence in the interaction. Additional supplementary information can also be accessed at http://prosecco.ucsd.edu/PanGIA/. Table 2.1 lists several public databases where protein and genetic interaction data can be downloaded for many different species.

Additional data files:

The file CYC2008_yeast_complexes.txt (Supplementary Data 3) contains a list of 408 protein complexes in the yeast S. cerevisiae hosted by the CYC2008 database[67,68]. This file illustrates an example of a Cytoscape node attribute file, which allows nodes in a network to be mapped to a particular attribute (Box 2). In this case, yeast genes are mapped to the various physical

complexes in which they participate. This file is used to demonstrate how a set of known biological modules can be used to train PanGIA to identify more biologically meaningful modules and intermodule relationships (covered in the 'Training PanGIA' PROCEDURE subsection, Steps 27–29). Additionally, this file is used during the 'Module labeling' section of this protocol (Steps 30–32) to check if the identified modules correspond to known protein complexes. Table 2.2 outlines several different public databases from which an annotation set can be downloaded for a variety of species.

Chapter 2.4: Procedure

Chapter 2.4.1: Steps 1-18: Importing physical and genetic networks into Cytoscape

1. Start Cytoscape. If Cytoscape is not yet installed on your computer, instructions for downloading and installing the latest version can be found at http://www.cytoscape.org/documentation_users.html. Cytoscape can be started by navigating to the directory in which it was installed and executing the file cytoscape.bat (Windows users) or cytoscape.sh (Linux and Mac OS X users). Critical step: PanGIA requires Cytoscape version 2.8.0 or higher. If your current installation of Cytoscape does not meet this requirement, download and install the latest version from http://www.cytoscape.org/.
2. Next, install the required plug-ins by navigating to the Plug-ins menu and clicking on Manage Plug-ins.

3. Double-click on the Analysis folder located under the Available for Install folder and select the plug-in for PanGIA version 1.1 or later. Click Install. Accept the plug-in license agreement and then click Finish.
4. Repeat the above step with BiNGO29 version 2.42 or later (located in the Functional Enrichment folder), EnhancedSearch30 version 1.2 or later (located in the Analysis folder) and CyThesaurus version 1.2 or later (located in the Network and Attribute I/O Folder).
5. After installing the required plug-ins, start the PanGIA plug-in by navigating to the Plug-ins menu and selecting Module Finders right arrow PanGIA.
6. After PanGIA has started, the PanGIA console will appear (Figure 2.3). The console is divided into three main panels: the Physical Network panel, where details regarding the physical network will be entered; the Genetic Network panel, where details regarding the genetic network will be entered; and the Advanced Options panel, which can be expanded by clicking on the triangle located next to the word 'Advanced'. This panel contains multiple advanced options for tuning the module-finding process. Four additional areas of interest are the Cytoscape canvas, which displays network visualizations and may be initially blank; the Data Panel, which is used to display node, edge and network attribute data; the Toolbar, which contains numerous command buttons; and the Network Browser, which can be accessed by clicking on the tab titled 'Network' (Figure 2.3). The Network Browser provides a list of networks currently available along with the number of nodes and edges in each network.

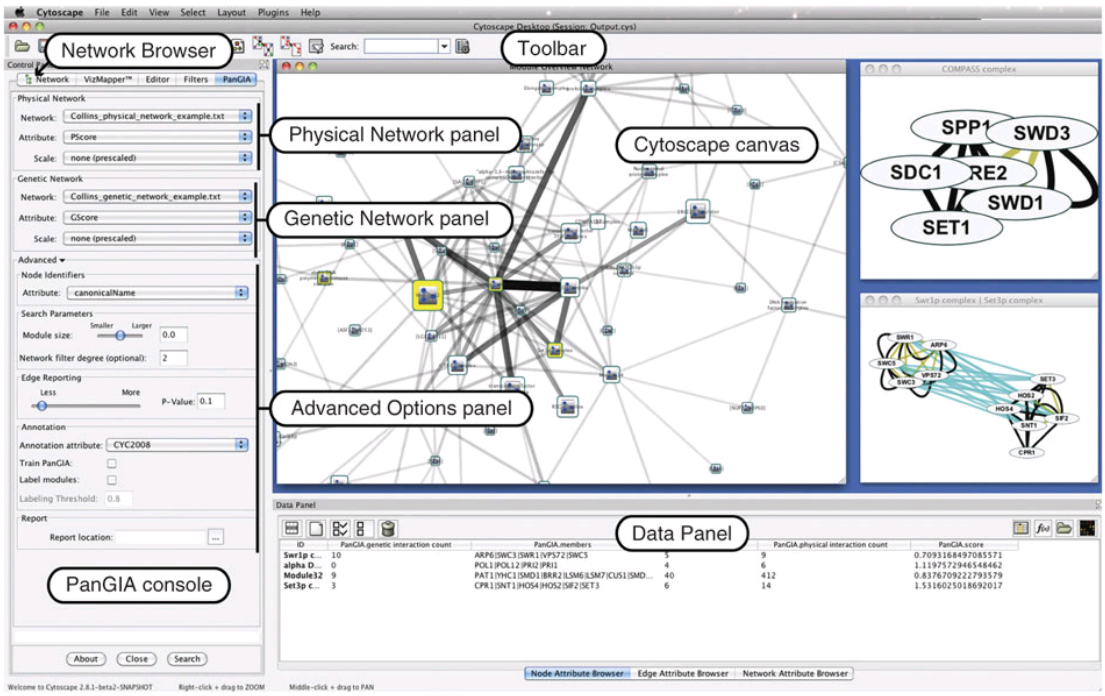


Figure 2.3: The PanGIA console.
The Cytoscape canvas displayed the network data and may initially be blank. The Data Panel (bottom) is used to display node, edge and network attribute data. The Toolbar (top) contains numerous command buttons used for navigating the network. The PanGIA console (left) is divided into three main panels, including the Physical Network panel, the Genetic Network panel and the Advanced Options panel. The Network Browser may be accessed by clicking on the Network tab located to the left of the PanGIA console tab.

7. Next, we import both a physical and a genetic network to be used in the analysis. Assemble the data in a tab-delimited format as described in Box 1. Users wishing to follow this protocol as a tutorial should download the Supplementary Data 1 (Collins_physical_network_example.txt) and Supplementary Data 2 (Collins_genetic_network_example.txt) and continue with Step 8. Critical step: PanGIA is designed to work with both quantitative and nonquantitative interaction data. However, any single network (either physical or genetic) must consist of a single type of interactions (i.e., either all quantitative interactions or all non-quantitative interactions).

8. Click on the File menu, then select Import right arrow Network from Table (Text/MS Excel). The Import Network and Edge Attributes from Table window will appear.
9. Click on the button titled 'Select File(s)' and specify the file containing the physical interaction network. A preview of the file should appear in the Preview panel located at the bottom. Select the column number representing the gene, which is the source node in the selection box titled 'Source Interaction'. Select the column number representing the target node in the Target Interaction selection box. If the example files (Supplementary Data 1 and Supplementary Data 2) are being used, the source and target nodes are, respectively, columns 1 and 2.
10. Specify an interaction type that will enable Cytoscape to differentiate between protein and genetic interactions. Check the box titled 'Show Text File Import Options' and, under Network Import Options, enter a meaningful string character in the Default Interaction box (e.g., 'pi' or 'gi', depending on whether physical interactions or genetic interactions are being imported).
11. Optional step: Use this step if quantitative interaction strengths are attached to the network. In the Preview panel launched in Step 9, left-click the column, which represents the quantitative attribute under the Preview panel, to enable the import of this attribute into Cytoscape. Right-click the same column, and, when prompted, type in an appropriate Attribute name (e.g., PScore or GScore, depending on whether the physical or genetic network is being imported); click OK. Make sure to note the name used. You will need it later when selecting the attribute to be used

in the training process. If the sample data are being used, the quantitative attribute for each interaction will be present in the third column. Critical step: The quantitative attribute provided should be either an integer (e.g., numbers such as 1, −2 or 514) or a floating point (e.g., numbers such as 2.343, −45.7687 or 74.3).

12. Click the Import button located in the lower right-hand corner. The physical network should now appear in the Cytoscape canvas area. The title of the network should be the name of the file provided.
13. Repeat Steps 8–12 to import the genetic network.
14. Optional step: Steps 14–18 should be used if the physical and genetic networks use different gene identifier systems (e.g., UniProt ID versus Ensembl ID). PanGIA requires that the two networks use the same gene identifier system. To convert between two gene identifier systems, assemble an ID translation file into a tab-delimited format as described in Box 3. This file should contain a map between the gene identifier system currently being used and the target gene identifier system. Users following this protocol as a tutorial using the sample data provided should skip to Step 19.
15. Optional step: Start the CyThesaurus plug-in by clicking on the Plug-ins menu and then selecting CyThesaurus. A window titled 'CyThesaurus plug-in' should appear.
16. Optional step: Configure the CyThesaurus plug-in to use the ID mapping file generated in Step 14 by clicking on ID Mapping Resources Configuration. A new window titled 'ID Mapping Source Configuration' will open up. In the left panel of this window, click on the folder titled 'Local Remote Files', which will bring up

another window titled 'File-based ID Mapping Resources Configuration'. Under the panel named 'Data source', click Select file to specify the location of the ID mapping file. Click on Open, then OK, and finally Close.

17. Optional step: Select both the physical and genetic networks by clicking on them in the Available Networks panel, then click the right arrow button. The two networks will appear in the Selected Networks panel.
18. Optional step: Choose the two different gene identifier names used in the genetic and physical network in the Source ID Type(s) selection box. In the Target ID Type selection box choose the target gene identifier you wish to map to. Finally, in the selection box titled 'All target ID(s) or first only?' select the option to keep the first target ID only. Next, click OK. A message will pop up indicating how many gene identifiers were successfully mapped.

Chapter 2.4.2: Steps 19-23: Generating a module map using the PanGIA plug-in: selecting the physical and genetic network

19. In the uppermost panel in the PanGIA console (Physical Network panel, see Figure 2.3), select the physical network to be used in the Network selection box. The name of the physical network will correspond to the name of the file from which the network was imported.
20. Select the genetic network to be used in the Network selection box located in the Genetic Network panel. Again, the name of the network will correspond to the name of the file from which it was imported.

21. Optional step: Use this step if quantitative interaction data are being used. In the Attribute drop-down menu located in the Physical Network panel, select the appropriate attribute name (i.e., the name assigned to the quantitative attribute for physical interactions from Step 11). Similarly, select the appropriate attribute name for genetic interactions in the Attribute drop-down menu located in the Genetic Network panel.
22. Optional step: Use this step if quantitative interaction data are being used and no biological annotation data are present. Even without a set of known complexes or pathways, PanGIA can leverage the confidence values assigned to each interaction (physical or genetic) to identify modules and intermodule links that contain highly confident interactions. However, it is necessary to let PanGIA know how the quantitative information is scaled. In the Scale selection menu located in both the Physical Network and Genetic Network subpanels (Figure 2.3), choose one of the following options: 'lower'—this option indicates that smaller quantitative values (both positive and negative) represent more confident interactions; 'upper'—this option indicates that larger quantitative values (both positive and negative) represent more confident interactions; or 'none (prescaled)'—this option should only be chosen if the quantitative attribute attached to either the physical or genetic interactions already represents the likelihood that a given interaction falls within a known biological module. This option enables the user to perform the training procedure outside of PanGIA and use the subsequent results in the module search process. If the example files are being used, simply choose 'none'. During the

training process, PanGIA will automatically scale the score attached to each interaction to reflect how likely that interaction is to fall either within a module or between two modules.

23. Optional step: Use this step if the gene identifiers in either the physical or genetic network were mapped to a new gene identifier. In the Advanced Options panel, select the target gene identifier to which genes in both networks were mapped to under the Node Identifiers subpanel. If no gene identifier mapping was performed or if the user is following this protocol with the sample data, skip to Step 24.

Chapter 2.4.3: Steps 24-26: Generating a module map using the PanGIA plug-in: setting the module size and edge reporting parameters (optional)

24. Optional step: PanGIA features a number of advanced options for tuning the search process. The size and number of modules returned by the search process can be controlled by changing the Module Size parameter (located in the Advanced Options panel). This can be done using the graphical slider in the Search Parameters panel. Dragging the slider to the right will result in fewer modules with larger average size, while dragging the slider to the left will result in more modules with a smaller average size (Figure 2.1b). The value of the Module Size parameter will be displayed in a text box to the right of the slider. It is recommended to leave the slider in its default position for the first run and to adjust it later if the results are unsatisfactory. For the sample data provided, set the Module Size parameter to −1.6 by moving the slider to the left.

25. Optional step: Often, the physical network being used covers a much larger set of proteins than those examined in the genetic interaction screen. In such a case, it is often useful to trim the physical network to include only proteins that are either present in the genetic network or are neighbors of such proteins within the physical network. This trimming is controlled by setting the 'network filter degree' parameter (located in the Advanced Options panel). A value of 0 will trim the physical network to only include nodes from the genetic network. Higher values represent the acceptable distance (through edges) separating a protein in the physical network from a node in the genetic network. If no trimming is desired, leave the box blank to prevent PanGIA from filtering any nodes. If the sample data file is being used, leave the network filter degree parameter at its default value of two. Critical step: The network filter degree parameter provided should be a positive integer (e.g., numbers such as 1, 2 or 10).
26. Optional step: Use this step only if quantitative interaction data are present. Every intermodule link found by PanGIA can be assigned a P value, after which insignificant edges are filtered from the resulting module map. The significance threshold can be set by changing the position of the slider in the Edge Reporting subpanel. Dragging the slider to the left (toward 'Less') will result in a higher significance threshold and less intermodule links in the final map (Figure 2.1c). The P value cutoff will be displayed in a text box immediately to the right of the slider. If the example files are being used, move the slider to the left and set the threshold to 0.05.

Chapter 2.4.4: Steps 27-29: Generating a module map using the PanGIA plug-in: training PanGIA (optional)

27. Optional step: Steps 27–29 should be used only if an annotation set is present. The training and module labeling steps require a list of annotations to be imported into Cytoscape. Assemble your list of annotations into the node attribute file format as described in Box 2. Import this file into Cytoscape by navigating to File right arrow Import right arrow Node Attribute.... Navigate to the appropriate file and click Open. If using the sample data, the file CYC2008_yeast_complexes.txt (Supplementary Data 3) should be used in this step.
28. Optional step: In the Annotation subpanel under Advanced Options, select the annotation attribute that will be used during the training and labeling process. The name of the annotation set is specified in the node attribute file, which was uploaded in the previous step (see Box 2 for more details). If the sample data have been used, the attribute name will be CYC2008. Select the annotation set name in the selection box titled Annotation attribute.
29. PanGIA can be trained to better identify module and intermodule links by examining actual examples of biological modules provided in the annotation set. To train PanGIA, simply check the box titled 'Train PanGIA' in the Annotation subpanel. If the sample data are being used, make sure this box is checked.

Chapter 2.4.5: Steps 30-32: Generating a module map using the PanGIA plug-in—labeling modules (optional)

30. Optional step: This step should only be used if an annotation set is present. PanGIA can label individual modules with the name of an annotation, if their member genes overlap with the genes belonging to that annotation (Figure 2.1c). To have PanGIA label modules, check the Label modules box in the Annotation subpanel (Figure 2.3). Next, specify the overlap threshold (defined here as the Jaccard index) in the Labeling Threshold text box. If the sample data are being used, set the Labeling Threshold to 0.2.
31. Optional step: If desired, PanGIA can output a report containing a summary of the module-finding process. This includes a summary of the networks used by PanGIA, the results of the training process and a summary of the resulting module map. To have PanGIA output a report, specify an output file in the Report subpanel. After a successful search, an HTML file will be created, which can be viewed using any Internet browser.
32. At this point, PanGIA is fully configured. The module search process can be initiated by clicking the Search button located at the bottom-right corner of the PanGIA console. Depending on the size of the network and the computer hardware, the module-finding process should take anywhere from 1 to 10 min. If the sample data are being used, the search process should take less than 1 min.

Chapter 2.4.6: Steps 33-35: Visualization of the module map using nested networks: navigating the module map

33. Once the search process is complete, a window titled 'Module Overview Network' will appear in the Cytoscape Canvas panel (Figure 2.4a). This network is the resulting global module map. Each node represents an individual module composed of a set of genes densely interconnected by genetic and physical interactions. The area of a module scales according to the number of genes that it contains. Links between modules are composed of genetic interactions; the thickness of the interactions corresponds to the number of genetic interactions spanning the two modules. If the labeling option was chosen, modules that overlap with one of the annotations provided will be labeled as such (Figure 2.4a,b).

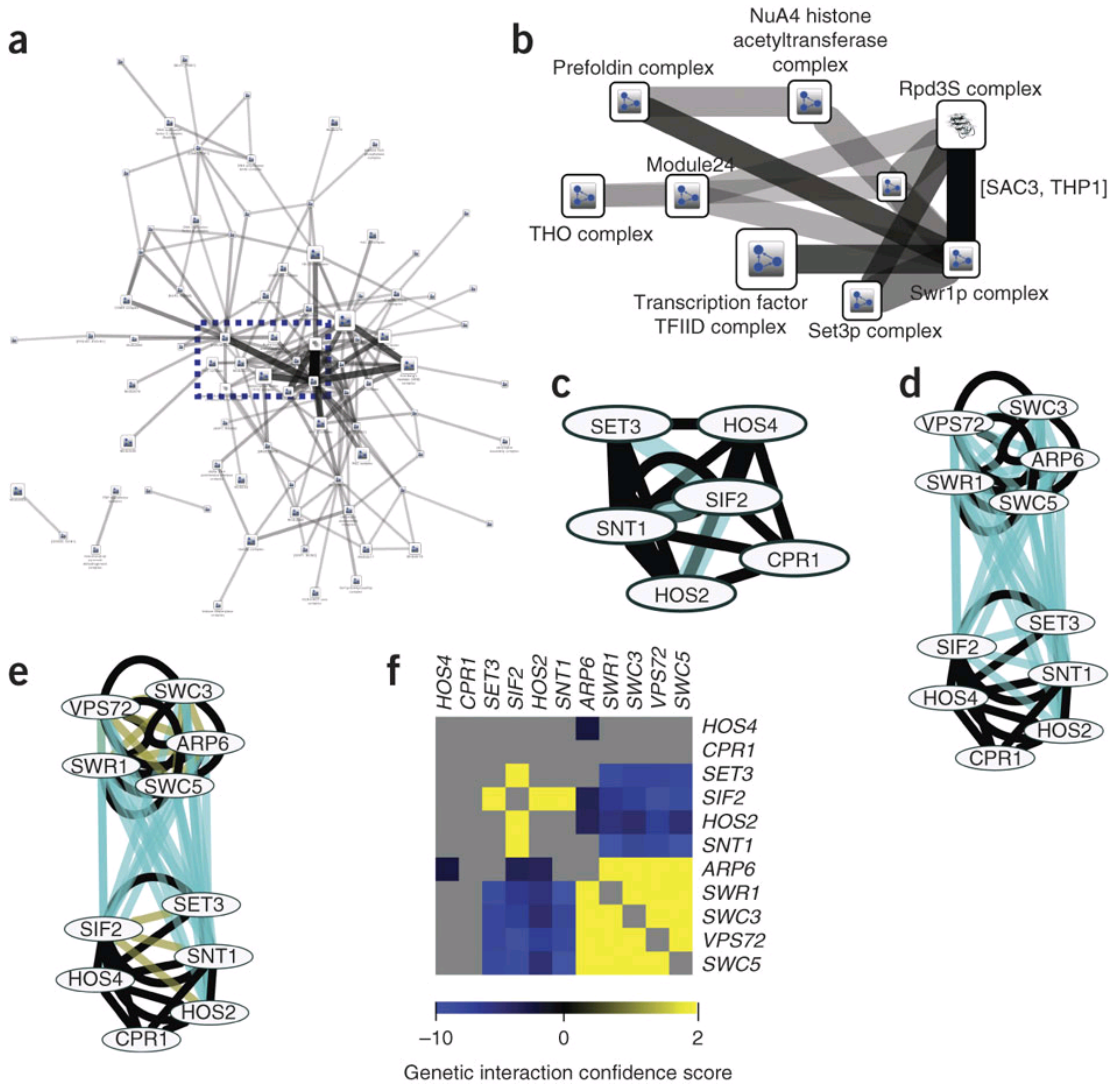


Figure 2.4: PanGIA output.

(**a**) The module map returned by PanGIA. Each node is a separate module or complex and the area of the node reflects the number of genes contained within the module. (**b**) A zoomed-in portion (blue box) of the module map shown in **a**. If an annotation set was provided and the labeling option was chosen, modules which overlap substantially with an annotation are labeled as such (e.g., Rpd3S complex). Modules not overlapping with any of the provided annotations are either given a generic name (e.g., Module 24) or labeled with a gene name (e.g., [*SAC3*,*THP1*]) if the module contains only one or two genes. (**c**) A detailed view of a single module. Each node represents a single gene that was assigned to this module. Physical interactions are colored black, whereas genetic interactions are colored turquoise. (**d**) A detailed view for two modules. Edges are colored similarly to **c**. The layout algorithm seeks to physically separate each module. (**e**) The same detailed view of two modules as shown in **d**, except that positive genetic interactions are colored yellow, whereas negative genetic interactions are colored turquoise. (**f**) The same network as shown in **e**, but visualized as a hierarchically clustered heat map using MeV[65].

34. You can zoom into the module map using the Zoom In button on the toolbar. This icon is displayed as a magnifying glass with a '+' symbol in the middle. You can zoom out by clicking on the Zoom Out button (magnifying glass with a '−' symbol in the middle). Alternatively, you can zoom in and out using the scroll wheel on the mouse. Scrolling up zooms into the area centered on the mouse pointer. Scrolling down zooms out on the area centered on the mouse pointer.
35. To pan around the module map, two options are available—using the mouse (option A) or using the network browser (option B):
    A. Using the mouse: Click the middle button on the mouse (or the scroll wheel, if present) anywhere in the active network being viewed in the Cytoscape canvas and drag the mouse in the desired direction.
    B. Using the network browser: Navigate to the 'Network Browser' by clicking on the Network tab (Figure 2.3) located to the left of the PanGIA tab. In the bottom half of the Network Browser is a bird's-eye view of the active network being viewed in the Cytoscape canvas; a blue selection box highlights the particular region of the network currently being viewed. To pan around the network, click and hold the blue selection box and move it in the desired direction.

Chapter 2.4.7: Step 36: Visualization of the module map using nested networks—identifying modules of interest

36. To further investigate modules of interest (i.e., function enrichment or detailed visualization), the module or modules of interest must be selected. We describe three different options for doing so: direct selection of modules (option A), direct selection of intermodule links (option B) and search-based selection of modules (option C).

    A. Direct selection of modules: Select any single module by clicking on it the with the left mouse button. The selected module will turn yellow. Several modules can be selected by holding down and dragging the left mouse button to define a rectangular selection region. Alternatively, multiple modules may be selected by holding down the shift button and left-clicking on multiple modules.

    B. Direct selection of intermodule links: To select any edge, click on the edge with the left mouse button. The selected edge will turn red. Several edges can be selected by holding down and dragging the left mouse button to define a rectangular selection region.

    C. Search-based selection of modules: To find and highlight modules in the map that contain a gene of interest, enter the name of the gene into the Enhanced Search plug-in search box located in the command toolbar

(Figure 2.3). If your gene of interest falls within a module, that module and its intermodule links will be highlighted yellow.

Chapter 2.4.8: Steps 37-45: Visualization of the module map using nested networks—exploring modules of interest

37. PanGIA returns numerous useful statistics or attributes regarding the modules identified, including module size, number of physical/genetic interactions among the genes in this module and so on. A complete list of attributes returned by PanGIA is provided in Table 2.3. The Data Panel (Figure 2.3) can display any/all of the attributes listed in Table 2.3. Select a module(s) of interest from the module map displayed in the Cytoscape Canvas as described in Step 36. When a single module or groups of modules have been selected in the Cytoscape Canvas, the selected modules will be listed in the Data Panel (Figure 2.3). Next, click on the Select Attributes button located in the upper left corner of the Data Panel. This will cause a list of attributes to appear; select which attributes you wish to view by clicking on their name. Exit this menu by clicking anywhere else.

Table 2.3: Description of Module-Level Attributes Returned by PanGIA.

Attribute Name	Attribute Type (Node or Edge)	Description
PanGIA Member Count	Node	Number of genes present in module
PanGIA Module Physical Interaction Count	Node	Number of physical interactions present in this module.
PanGIA Module Genetic Interaction Count	Node	Number of genetic interactions present in this module.
PanGIA Source Size	Edge	Member count of the source module
PanGIA Target Size	Edge	Member count of the target module
PanGIA Genetic Interaction Count	Edge	Number of genetic interactions spanning the two modules connected by this edge
PanGIA Physical Interaction Count	Edge	Number of physical interactions spanning the two modules connected by this edge
PanGIA P-value	Edge	Significance of the inter-module link
PanGIA Edge Score	Edge	The total score of genetic interactions spanning two modules minus the score of the physical interactions
PanGIA Genetic Interaction Density	Edge	Represents the Edge Score divided by the Genetic Interaction Count.

38. The Data Panel can also display detailed information regarding intermodule links in the map. Select one or more intermodule links of interest in the map as described in Step 36. In the Data Panel, click on the tab labeled Edge Attribute Browser. The panel will display the edges that have been selected. Similar to the modules, intermodule links identified by PanGIA also have several informative attributes as outlined in Table 2.3. These attributes can be viewed by selecting them through the Select Attributes menu (see Step 37).

39. To visually inspect a single module or a group of modules in greater detail, select the module(s) of interest as outlined in Step 36. Next, right-click any of the selected module(s) and choose PanGIA right arrow Create Detailed View. A new window will appear in the Cytoscape Canvas area containing the module (Figure 2.4c) or modules (Figure 2.4d) of interest. In this detailed view, each node represents a single gene. Edges represent either physical interactions (colored black) or genetic interactions (colored turquoise). If quantitative genetic interaction data are used, positive genetic interactions will be colored yellow, whereas negative genetic interactions will be colored turquoise (Figure 2.4e).
40. The network displayed in the detailed view can be laid out and manipulated similarly to the module map as described in Steps 33–35. Individual genes and interactions between genes can be selected similarly to the way in which modules are selected in the module map as described in Step 36.
41. Optional step: Steps 41–44 should be followed if quantitative interaction data are present. An alternate means of visualizing a single module or a set of connected modules is via a hierarchically clustered heat map (Figure 2.4f). In this view, each row or column represents a single gene. Each cell in the matrix is colored to represent the quantitative value attached to the interaction between those two genes. For example, Figure 2.4f is a hierarchically clustered representation of the between-cluster model shown in Figure 2.4e. The colors in the heat map represent the genetic interaction confidence scores between the genes. PanGIA can output a matrix containing either the genetic interaction confidence scores or physical

interaction confidence scores between individual genes (option A), between all genes in a module or set of modules (option B):

A. Output interaction matrix for a select number of genes
    a. Select the genes of interest from a detailed view as described in Step 36. Right-click on any of the selected genes and select PanGIA right arrow Save Selected Nodes to Matrix File.
    b. Next, choose the desired quantitative attribute to be outputted (i.e., physical interaction confidence or genetic interaction confidence). The names of these quantitative attributes will be the ones assigned by the user in Step 11.
    c. A dialog box will appear prompting to you enter the output file name. Enter the file name and click Save.
B. Output interaction matrix for all genes in a module or set of modules
    a. Select a module(s) of interest as outlined in Step 36. Right-click on any of the selected modules and select PanGIA right arrow Save Selected Nodes to Matrix File.
    b. Choose the desired attribute to be outputted. Enter the output filename and click Save. If you are using the Sample data, select the modules labeled 'Swr1p complex' and 'Set3p complex'. Right-click on one of these two modules and select PanGIA right arrow Save Selected Nodes to Matrix File right arrow GScore.
    c. Provide an appropriate file name and click Save.

42. Optional step: Start the MeV program. The Multiple Array Viewer window should pop up. Load the interaction matrix generated in the previous step by navigating to File right arrow Load Data. The Expression File Loader dialog window will appear. Click the 'Browse' button and specify the file containing the interaction matrix. A preview of the interaction matrix should appear in the Expression Table panel. Click the upper-leftmost interaction confidence score and then click Load. A heat map of the interaction matrix will appear in the Multiple Array Viewer window.
43. Optional step: To hierarchically cluster the heat map, click on the Clustering tab located near the top of the window and then select Hierarchical Clustering. In the HCL: Hierarchical Clustering window that will open, check the boxes to Optimize Gene Leaf Order and Optimize Sample Leaf Order. This will ensure that genes with similar interaction profiles will be placed close to one another. Finally, click OK.
44. Optional step: In the rightmost panel of the Multiple Array Viewer navigate to Analysis Results right arrow HCL (1) right arrow HCL Tree. A hierarchically clustered version of the heat map will appear. This image can be saved by clicking on File right arrow Save Image. Multiple output formats are available. If using the example data, the heat map should look similar to Figure 2.4f.
45. In cases in which a module may contain one or more genes with an unknown function, it is useful to be able to query an external web-based database such as Ensembl or Entrez. Cytoscape features the ability to automatically connect to and

query external web databases. Right-click on a gene of interest within the Detailed View and navigate to the LinkOut menu. Numerous databases will be listed including Ensembl, KEGG, UniProt and Entrez. Select one of these databases. An Internet browser window will open automatically displaying any information the selected database has on the gene of interest. This feature provides an effective way to interrogate the function of unannotated genes.

Chapter 2.4.9: Steps 46-49: Functional enrichment of the modules

46. Start the BiNGO plug-in by selecting Plug-ins right arrow Start BiNGO. The BiNGO Settings window will appear.
47. Select the module or modules of interest that will be examined for an enriched function. Create a Detailed View as outlined in Step 39. Select the genes contained in the module(s) that will be screened for an enriched GO function. To select all genes, simply press Ctrl + A simultaneously (or Command + A, if using Mac OS X).
48. Type in a meaningful name for the set of genes being examined in the box titled 'Cluster name'. Under the Select Organism/Annotation menu, choose the appropriate organism (for the sample data choose Saccharomyces cerevisiae). For the remaining options, the default values will typically suffice. Click Start BiNGO. Depending on the number of genes selected and the computer hardware, this process will take 5–10 min.

49. BiNGO will return an output window containing a list of GO terms that were found to be enriched along with their respective P values. BiNGO will also return a network of GO terms showing the inter-relationships between the various GO terms that were found to be enriched. The color of each term represents its significance of enrichment.

Chapter 2.4.10: Step 50: Exporting your results

50. Cytoscape enables multiple ways to export individual modules as well as the global module map. For a thorough explanation of each of these export methods, please refer to the online tutorial (http://www.cytoscape.org/documentation_users.html). Note: for general troubleshooting and timing advice, please refer to Tables 2.4 and 2.5.
    A. Export network as a graphics object
        a. The module map, as well as individual modules, can be exported as a graphics file. Numerous output formats are supported including PDF, JPEG, SVG, PNG and BMP.
        b. To export a network as a graphics object, make sure it is the active window and then select File right arrow Export right arrow Network View as Graphics....
        c. In the Export Network View as Graphics dialog box, select the output file name and choose the desired output format. Click OK.

d. If the graphics object will be further manipulated in a graphics software package, such as Adobe Illustrator, we recommend exporting the network as a PDF file. Make sure to also check the box titled 'Export text as font', which will enable the manipulation of the text labels in the network image.

B. Export modules as a tab-delimited file

a. Each of the individual modules can be exported in a tab-delimited file, where each line consists of two parts separated by a tab character: the name of the module and the genes comprising the module. If multiple genes have been assigned to a module, each gene will be separated by the '|' character.

b. To export the modules as a tab-delimited file, right-click on any module in the module map (i.e., the Module Overview Network in the Cytoscape Canvas) and select PanGIA right arrow Export right arrow Export Modules to Tab-Delimited file.

c. Specify the output file in the dialog box that pops up and click Save.

C. Export module map as a tab-delimited file

a. The entire module map can be exported as a tab-delimited file, where each single line represents a single interaction between two modules. A single line is split into nine different parts separated by a tab character. The first two parts represent the source and target module. The

remaining seven parts represent various attributes describing each interaction as outlined in Table 2.3.

b. To export the module map as a tab-delimited file, right-click on any module in the module map (i.e., the Module Overview Network in the Cytoscape Canvas) and select PanGIA right arrow Export right arrow Export Module Map to Tab-Delimited file.

c. Specify the output file and click Save.

D. Export the entire PanGIA session as a Cytoscape session file

a. The entire PanGIA session can be saved to file. A session file contains all of the results of this entire workflow. This includes all networks that were loaded or generated (physical, genetic, module map, individual modules), any custom visualization styles that were employed and any enrichment results obtained from BiNGO. Saving to a session file will enable the user to continue the analysis at a later point.

b. To save the entire PanGIA session to file, select File right arrow Save As. Type in the name of the output file and click Save.

Chapter 2.5: Troubleshooting

Troubleshooting advice for specific steps in the protocol can be found in Table 2.4. In addition, we outline two of the biggest problems a user may face and potential solutions to these problems below:

Table 2.4: Troubleshooting Table

Step	Problem	Possible Reason	Solution
1	Executing cytoscape.bat (Windows) or Cytoscape.sh (Mac OSX, Linux) does not open Cytoscape.	Java is not installed properly.	Make sure Java Version 1.6.014 or higher is installed. Java can be downloaded at http://www.java.com
21	PanGIA fails to label any of the modules in the final module map.	Threshold for labeling may be set too high.	Set the labeling threshold slightly lower to allow more modules to be labeled.
23	The module search process is taking a very long time.	Insufficient memory and/or processing power.	Very large physical or genetic networks (>500,000 interactions) require a larger amount of memory than specified in the Equipment Setup section. See the Timing section for recommendations on the amount of memory and processing power required for larger networks.
34	The queried database fails to return any information on the selected gene(s) of interest.	Mismatched gene identifiers.	When querying an external database, the identifier of the selected gene(s) must be identical to the identifier used by the external database. For example, if querying the Ensemble database, selected genes need to use Ensembl identifiers in order to have any information returned. Use one of the recommended website to map gene identifiers if there is any discrepancy[26,27].
37	BiNGO supplies an error message asking to 'Please select one or more nodes.'	No genes were selected for examining functional enrichment.	Visualize the module(s) of interest as outlined in Step 27. In the detailed view, select one or more genes of interest. All nodes (genes) can be selected in a detailed view by pressing 'Ctrl' (or 'Cmd' if using Mac OS X) + 'A'

Module size issues: In some cases PanGIA may fail to return any modules or it may return modules that are either very large or very small (i.e., that consist of a single gene). The problem may be addressed by moving the Module Size slider bar in the Advanced Options panel (see Step 24). Dragging the slider to the right will generally result in fewer but larger modules. Dragging it to the left will have the opposite effect. Once the slider has been set to a new position, make sure the rest of PanGIA is

properly configured (Steps 19-31) and hit the Search button located at the bottom of the PanGIA console.

Edge reporting issues: Another common issue is that the module map may contain either too few or too many intermodule links. PanGIA utilizes a sampling-based procedure to assign P values to every intermodule link and only those links with a P value below a specified threshold are displayed in the final module map. If the threshold is set too high, this may cause a number of spurious interactions to appear in the module. On the other hand, if the threshold is set too low, this may cause PanGIA to filter out intermodule links of biological interest. This problem may be addressed by adjusting the threshold by moving the Edge Reporting slider bar in the Advanced panel (as described in Step 26). Moving the slider to the right will result in a higher threshold and subsequently a larger number of intermodule links in the final map. Moving it to the left will have the opposite effect.

Chapter 2.6: Timing

The time required to complete this protocol is almost entirely dependent on the size of the genetic and physical networks being analyzed. Table 2.5 charts the amount of time required for the module-search process (under default options) using networks of various sizes as input. For a physical and genetic network containing less than 100,000 interactions each (~200,000 interactions total), PanGIA takes, on average, ~10 min.

Table 2.5: Time Required to Run PanGIA on Networks of Various Sizes

Number of interactions (Genetic + Physical)	Run Time	
	Processor: Dual Core, 32-bit (3.2 GHz) Memory: 2 gb Graphics Card Memory: 256 mb	Processor: 8-core, 64-bit (2.8 GHz) Memory: 8 gb Graphics Card Memory: 256 mb
10,000	<1 minute	<30 seconds
50,000	1 minute	<1 minute
100,000	2 minutes	1.5 minutes
500,000	15 minutes	10 minutes
1,000,000	Insufficient Memory	30 minutes

Chapter 2.7: Anticipated results

Using the sample physical (Supplementary Data 1) and genetic (Supplementary Data 2) interaction networks with PanGIA, configured as suggested in this protocol (module size parameter = −1.6, edge filtering parameter = 0.05, network filter = 2, training enabled, labeling threshold = 0.2), will produce a module map containing 82 modules and 164 intermodule links (Figure 2.4a). Overall, 34 of these modules overlap with known complexes provided in the file CYC2008_yeast_complexes.txt (Supplementary Data 3) and will be labeled accordingly.

The resulting module map provides a wealth of hypotheses that can form the basis for follow-up experiments. Because PanGIA has been trained on databases of known complexes and pathways, it is likely that many modules will correspond to known protein complexes in the PanGIA results[19,30,58]. Other modules that do not correspond to prior knowledge are prime candidates for novel complexes or pathways. The module map produced using the sample data contains 21 modules (out of 82) with

two or more genes that do not overlap with any known S. cerevisiae physical complexes. One could test the members of these 21 modules for co-complex membership. An alternate strategy for revealing novel biological functions is to identify modules that are enriched for a common biological function, yet contain some genes that are not yet annotated to that particular function. For example, Module 24 (Figure 2.4b) is enriched for genes involved in nuclear pore organization ($P < 7.05 \times 10^{-11}$). However, two of the genes in Module 24, SEC31 and SEC16, are not annotated to this function. The logical hypothesis in this case would be that these two genes are involved in nuclear pore organization and that a deletion or knockdown of these genes should have an impact on this function.

Intermodule links, on the other hand, predict functional overlap or synergy between the two connected modules[19,58]. For example, a large number of genetic interactions span the two modules corresponding to the Rpd3S complex and Swr1p complex (Figure 2.4d,e). The Swr1p complex has been well established as a chromatin remodeler, which deposits H2A.Z, a histone variant, onto chromatin. The function of the Set3p complex is less well understood. The intermodule link between the two complexes suggests that Set3p may have a role similar to that of the Swr1p complex. Indeed, a recent publication has provided evidence suggesting that this may be the case[69].

Chapter 2.8: Acknowledgments

Srivas R, Hannum G, Ruscheinski J, Ono K, Wang PL, Smoot M, Ideker T. (2011) Assembling global maps of cellular function through integrative analysis of physical and genetic networks. Nature Protocols 6,1308–1323 (2011) doi:10.1038/nprot.2011.368.

Gregory and Rohith contributed equally to this work. Contributions: G.H., R.S. and T.I. conceived and led the project. G.H. coded PanGIA with supporting code from R.S., J.R., K.O., P.-L.W. and M.S. R.S., G.H. and T.I. wrote the paper. All authors have contributed to the design of PanGIA and all have read and approved the paper.

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

We gratefully acknowledge S. Bandyophadyay and R. Kelley for their role in the development of the framework used in PanGIA. M. Michaut provided useful feedback on the manuscript. C. Doherty and M. Ashkenazi provided helpful beta testing of the PanGIA plug-in. This study was supported by grants from the National Institute of General Medical Sciences (GM070743), the National Science Foundation (IIS0803937) and Microsoft (Computational Challenges in Genome-wide Association Studies).

CHAPTER 3: GENOME-WIDE METHYLATION PROFILES REVEAL QUANTITATIVE VIEWS OF HUMAN AGEING RATES

Chapter 3.1: Summary

The ability to measure human ageing from molecular profiles has practical implications in many fields, including disease prevention and treatment, forensics, and extension of life. In particular, ageing has been linked to changes in DNA methylation, but the nature and extent of this relationship are not well understood. Here, we investigate the changes in genome-wide methylation patterns as people age, using quantitative measurements at more than 450,000 CpG markers from the whole blood of 679 human individuals, aged 19 to 101. We find two distinct signatures of ageing in the data and show that these signatures are reflected in the transcriptome. Building on these results, we formulate a predictive model for the rate at which an individual's methylome ages, and we show that this rate is impacted by gender and genetic variants. Our ageing model highlights specific components of the ageing process and provides a quantitative read-out for studying the role of methylation in age-related disease.

Chapter 3.2: Introduction

Not everyone ages in the same manner. It is well known that women tend to live longer than men, and lifestyle choices such as smoking and physical fitness can hasten or delay the ageing process[70,71]. These observations have led to the search for molecular markers of age which can be used to predict, monitor, and provide insight into age-associated physiological decline and disease. One such marker is telomere

length, a molecular trait strongly correlated with age[72] which has been shown to have an accelerated rate of decay under environmental stress[73,74]. Another marker is gene expression, especially for genes that function in metabolic and DNA repair pathways which are predictive of age across a range of different tissue types and organisms[75–77].

A growing body of research has reported associations between age and the state of the epigenome— the set of modifications to DNA other than changes in the primary nucleotide sequence[78]. In particular, DNA methylation associates with chronological age over long time scales[79–83] and changes in methylation have been linked to complex age-associated diseases such as metabolic disease[84] and cancer[85,86]. Studies have also observed a phenomenon dubbed "epigenetic drift", whereby the DNA methylation marks in identical twins increasingly differ as a function of age[81,87]. Thus, the idea of the epigenome as a fixed imprint is giving way to the model of the epigenome as a dynamic landscape that reflects a variety of chronological changes. The current challenge is to determine whether these changes can be systematically described and modeled to detect different rates of human ageing, and to tie these rates to related clinical or environmental factors.

The mechanisms that drive changes in the ageing methylome are not well understood, although they have been attributed to at least two underlying factors[87,88]. First, it is possible that environmental exposure will over time activate cellular programs associated with consistent and predictable changes in the epigenome. For example, stress has been shown to alter gene expression patterns through specific changes in DNA methylation[89]. Alternatively, spontaneous epigenetic changes may

occur with or without environmental stress, leading to fundamentally unpredictable differences in the epigenome between ageing individuals. Spontaneous changes may be caused by chemical agents that disrupt DNA methyl groups or through errors in copying methylation states during DNA replication. Both mechanisms lead to differences between the methylomes of ageing individuals, suggesting that quantitative measurements of methylome states may identify factors involved with slowed or accelerated rates of ageing.

To better understand how the methylome ages and to determine whether human ageing rates can be quantified, we initiated a project to perform genome-wide methylomic profiling of a large cohort of individuals spanning a wide age range. Analysis of these data reveals two distinct signatures of ageing and suggests that age-associated changes in the methylome lead to changes in transcriptional patterns over time. Based on these findings, we construct a predictive model of ageing rate which we show is influenced by gender and specific genetic variants. These findings are replicated in a second large cohort.

Chapter 3.3: Results

Chapter 3.3.1 Global data on the ageing methylome

We obtained methylome-wide profiles for a mixed population of 397 Caucasian and 91 Hispanic individuals. Samples were taken as whole blood and processed using the Illumina Infinium HumanMethylation450 BeadChip[90] assay, which measures the methylation states of 485,577 CpG markers. Methylation was

recorded as a fraction between zero and one, representing the frequency of methylation of a given CpG marker across the population of blood cells taken from a single individual. Conservative quality controls were applied to filter spurious markers and samples (Methods). The distribution of methylation was bimodal, with most markers methylated at either very high or very low levels across all ages (Figures 3.1a,b), which is consistent with the findings of previous studies of the methylome[91]. The resulting data set represents the largest and highest-resolution collection of methylation data produced for the study of ageing, providing an unprecedented opportunity to understand the role of epigenetics in the ageing process. The complete methylation profiles are available at the Sage Bionetworks Commons (http://www.sagebase.org/commons/repository.php).

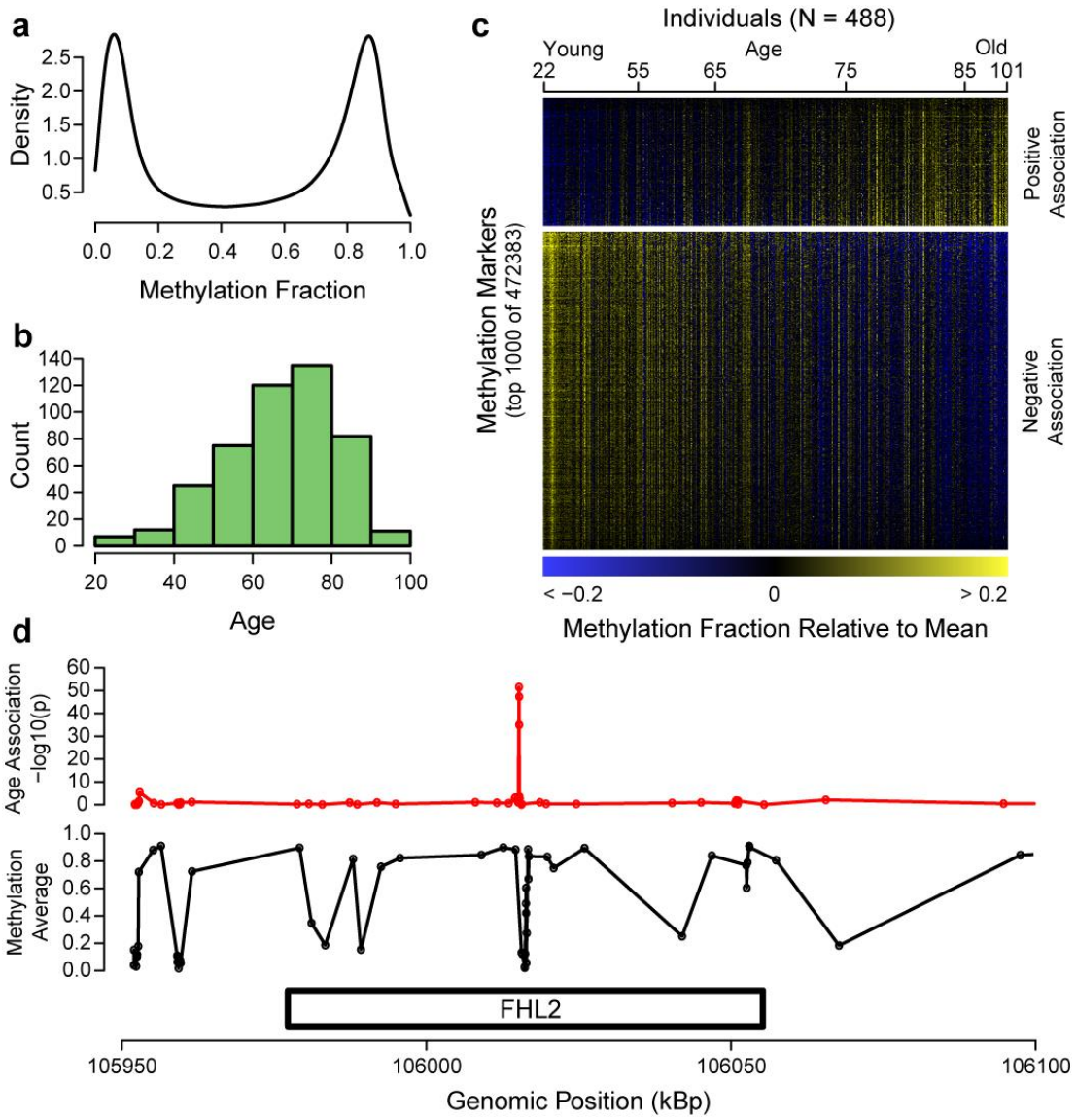


Figure 3.1 A high-density methylation map of human ageing

a, A density plot of methylation fraction values across all markers and individuals. Markers tend to be either predominantly methylated or not. **b,** A histogram of the age distribution for all individuals. **c,** A heat map of the top 1000 age-associated methylation markers, sorted by the magnitude of association (regression coefficient). The individuals are ordered youngest to oldest. **d,** An example association map for the gene FHL2. A strong ageing association is shown for several markers (red: -log10(p-value)) at a CpG island in the center of the gene (black: average methylation fraction).

Chapter 3.3.2 Two signatures of ageing

We were able to identify a robust ageing signal in the methylome data at high resolution, with at least 65,473 (14%) of the markers having significant associations between methylation fraction and age (Figure 3.1c, FDR < 0.05 by F-Test). One example is the gene Four and a Half LIM Domains 2 (FHL2), in which several strong age-associated methylation markers were identified within a single CpG island (an area enriched for CpG sites), coincident with an internal promoter (Figure 3.1d). This finding sheds light on previous work tying FHL2 to ageing and tumorigenesis[92,93]. Genes with nearby age-associated markers were enriched for functions in epigenetic regulation, such as sequence-specific DNA binding and cell differentiation (Methods, Supplementary Table 1). Interestingly, these genes were depleted for key DNA replication functions, including cell cycle arrest and DNA repair, suggesting that these essential functions are less tolerant to, or protected from, ageing-related changes in methylation. As a positive controls on this analysis, we observed that the age-associated markers included most CpG sites found to be associated with age in two previous studies which surveyed approximately 5% of those covered by the present data set[82,83] (significance of overlap $P < 10^{-98}$, $P < 10^{-34}$, by Fisher exact test).

In addition to age-associated changes in methylation fraction, we also observed evidence for a second ageing signature: many methylation markers become less stable with age, such that their variance in methylation fraction is greater among older individuals than younger individuals. To quantify this effect, we computed the deviance of each marker value as its squared-distance from the expected population

mean (Figure 3.2a, Methods). Then, in addition to testing for markers whose methylation fraction changes with age (Figures 3.2b,c), we were able to test for markers whose deviance changes with age (Figures 3.2d,e)[94]. Increasing deviance was a widespread phenomenon—we identified 10,083 markers for which the deviance was significantly associated with age (FDR < 0.05), of which 10,056 (99.7%) represented increased rather than decreased deviance (Figure 3.2e). While there was some concordance between the two signatures of ageing (association with methylation fraction versus association with methylation deviance, Spearman R = 0.31), the changes in methylation fraction were far more numerous and there were 2996 markers with an increase in deviance with no change in mean (Supplementary Figure 1). Thus, these two signatures— in which the methylation fraction and/or deviance associate with age— can act independently, raising the question of whether they reflect one or more than one ageing process.

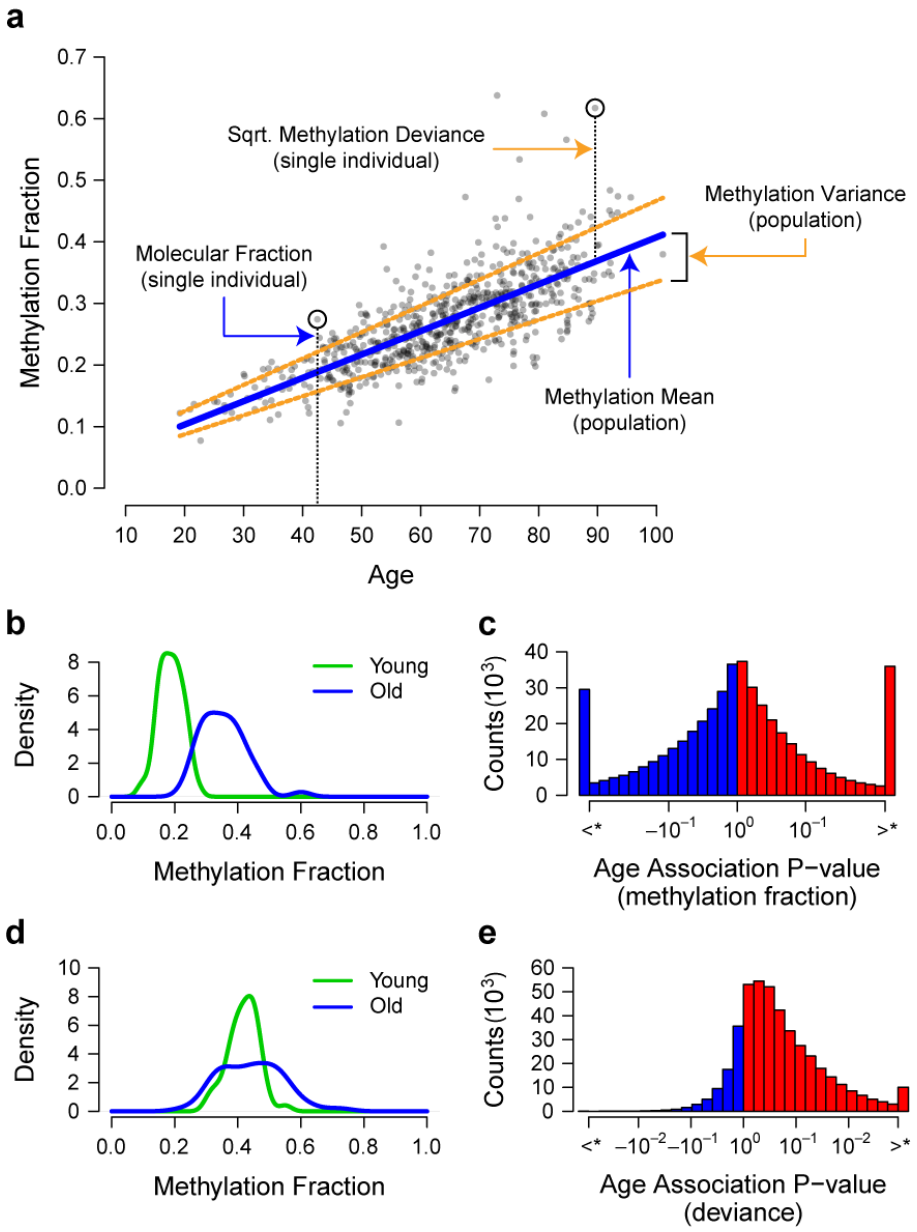


Figure 3.2 Methylation marker trends with age

a, Methylation fraction values for are shown for the marker cg24724428. Over any subset of the cohort, we consider two group methylation statistics: the mean and variance. Marker variance is a measure of the mean methylation deviance, which is defined as the squared difference between an individual's methylation fraction and their expected methylation fraction. **b,** A density plot showing the change in mean methylation with age for the marker cg24724428. Young and old groups are based on the top and bottom 10%. **c,** A histogram of the significance of association between the methylation fraction of all markers and age. P-values are signed such that positive values represent an increase of methylation with age. Markers which exceeded the FDR < 0.05 threshold are grouped into the most extreme bins. **d,** A density plot showing the change in methylation deviance with age for the marker cg24724428. **e,** A histogram in the same form as 'd', of the significance of association between the methylation deviance of all markers and age.

One way to conceptualize the two ageing signatures is in terms of Shannon Entropy, or loss of information content in the human methylome over time[95]. An increase in entropy of a CpG marker means that its methylation state in the cell population becomes less predictable over time, i.e. its methylation fraction tends towards more moderate values (Methods). Indeed, over all markers associated with a change in methylation fraction in the sample cohort, 81% tended towards a methylation fraction of 50% (Figure 3.3a, Binomial P ~ 0, Supplementary Table 2). The corresponding increase in methylome entropy with age was highly significant ($R = 0.17$, $p = 1.3 \times 10^{-4}$, Figure 3.3b). This suggests that the ageing methylome is characterized by increased heterogeneity across cells, possibly due to stochastic changes to the methylome or the composition of the cell population.

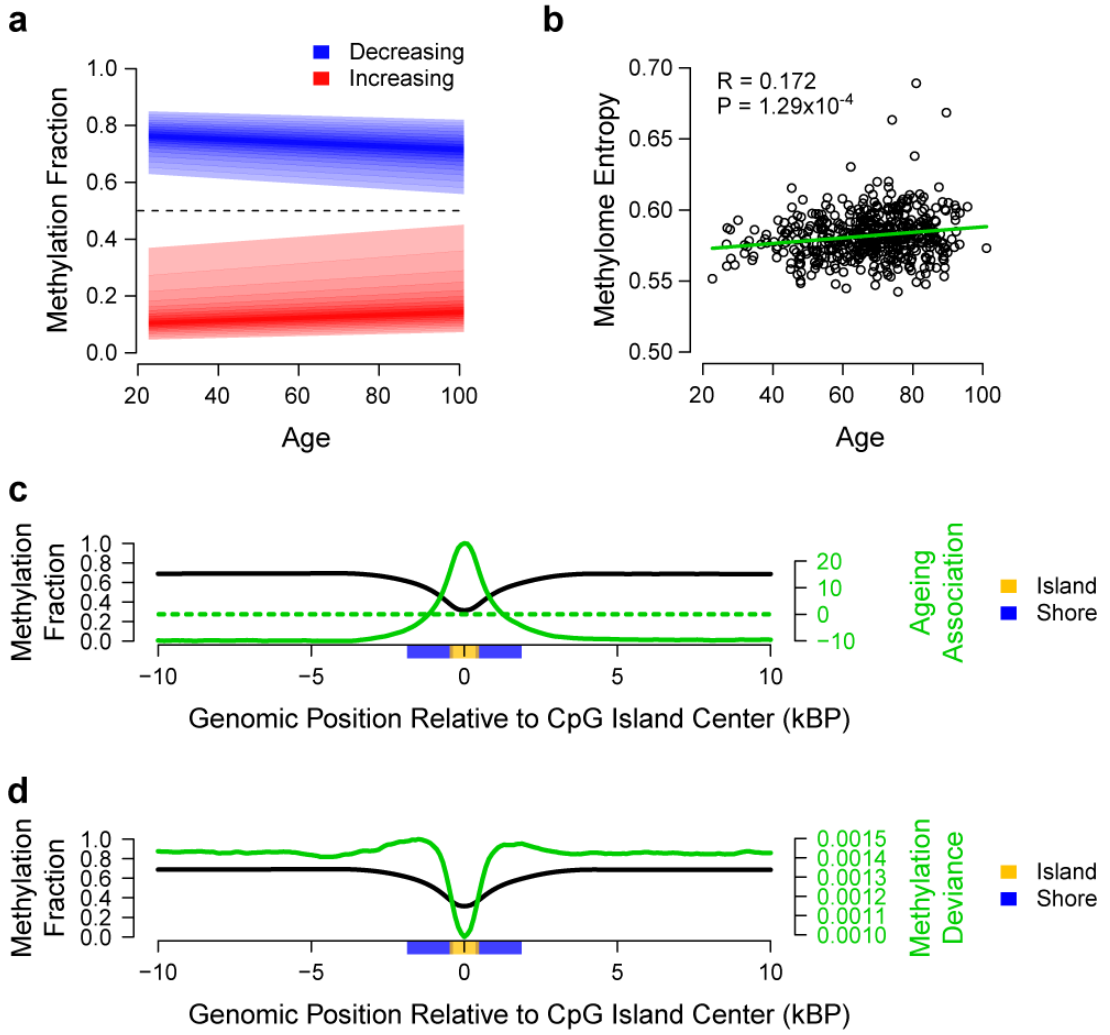


Figure 3.3 Methylome-wide trends with age

a, All methylation fraction values for increasing markers (red) and decreasing methylation markers (green) were regressed against age. The darkest colour represents the median and the bounds represent the 25% and 75% quantile. Both tend to converge to moderate methylation fraction values. **b,** Corresponding to 50% methylation, methylome entropy increases with age. **c,** An aggregate genomic map of the methylation fraction for 27,176 CpG islands (black). The ageing coefficient relating methylation fraction to age is shown in the same region (green). Color bars indicating the island and shore regions represent 75% confidence intervals. **d,** The same CpG island map (black), shown with the average methylation deviance in the same region (green).

Chapter 3.3.3 Correspondence with the transcriptome

As changes in methylation have been directly linked to changes in gene expression[96], we were interested in whether the two ageing signatures were mirrored

in the human transcriptome. For this purpose, we obtained and analyzed publicly-available gene expression profiles from the whole blood of 488 individuals spanning an age range of 20 to 75[97]. We found strong evidence for both signatures: i.e., genes whose expression associates with age (412 genes, FDR < 0.05) and increasing expression deviance (Binomial $P < 10^{-276}$, Methods). Furthermore, genes with age-associated expression profiles were also more likely to have nearby age-associated methylation markers ($P < 0.01$, Supplementary Table 3). Thus, the age-associated gene expression patterns are at least partially explained by age-associated changes in the methylome. These findings are supported by previous studies within ageing mouse populations, which reported a loss of molecular coordination with age in murine gene expression data[98,99].

A link between the transcriptome and the methylome was further supported by the incidence of age-associated methylation markers within CpG islands. CpG islands are often coincident with gene promoters and are un-methylated to permit gene expression; increases in methylation of these regions have been associated with ageing and cancer[79,100]. Indeed, we found that CpG islands mapped to highly significant increases in methylation fraction with age (Figure 3.3c, Supplementary Table 2). Moreover, we found that the methylation states of CpG islands were highly stable across individuals, as indicated by a sharp reduction in methylation deviance compared to other regions of the methylome (Figure 3.3d). These findings indicate that the methylation states of CpG islands are perhaps more tightly regulated than has been appreciated, such that very small changes to CpG islands are indicative of

advanced age. We further investigated the boundaries of CpG islands, known as shores. Contrary to the CpG islands, we observed a 23% increase of deviance in the shores (Figure 3.3d). This supports recent literature that describes the shores as active regulatory elements[101], which may be more susceptible to variation between individuals.

Chapter 3.3.4 A predictive model for the ageing methylome

Given the reproducible signal of age present in the methylome, we used a penalized multivariate regression model[102] and bootstrap (Methods) to build a predictive model of ageing that included both methylomic and clinical parameters such as gender and Body Mass Index (BMI) (Figure 3.4a). The optimal model selected a set of 67 methylation markers which were highly predictive of age (Figure 3.4a, Supplementary Table 4). The accuracy of the model was high, with a correlation between age and predicted age of 96% and an error of 3.8 years (Figure 3.4b). Nearly all markers in the model lay within or near genes with known functions in ageing-related conditions including Alzheimer's disease, cancer, tissue degradation, DNA damage, and oxidative stress. By way of example, two markers lay within the gene somatostatin (SST), a key regulator of endocrine and nervous system function[103]. SST is known to decline with age and has been linked to Alzheimer's disease[104]. As a second example, five model markers lay within the transcription factor KLF14, which has been called a 'master regulator' of obesity and other metabolic traits[105]. Given the

links between ageing, longevity, and metabolic activity[106,107], it is not surprising that several of our model markers are implicated in obesity and metabolism.

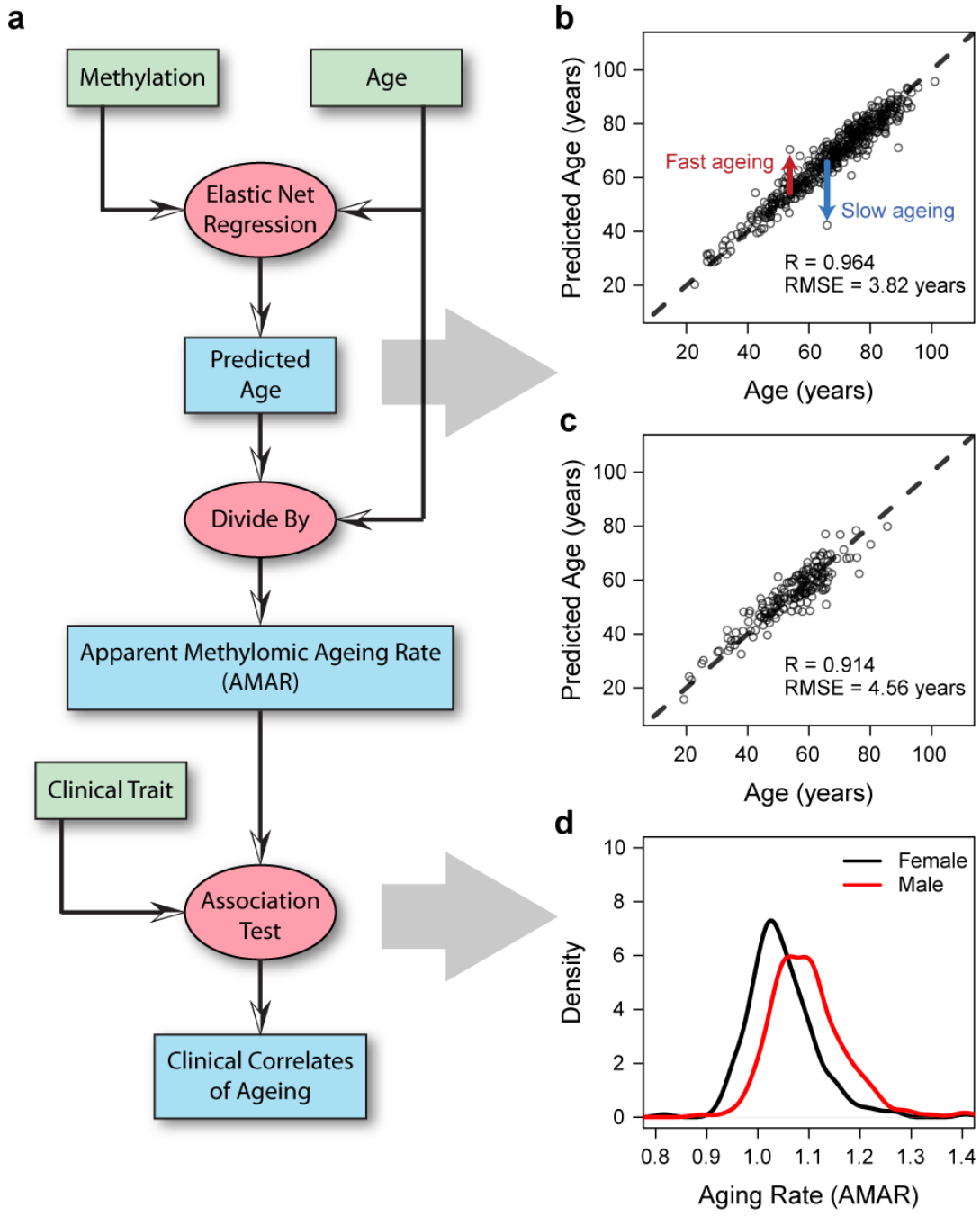


Figure 3.4 Model predictions and clinical variables

a, A flow chart of the data (green boxes) and analyses (red ovals) used to generate ageing predictions (blue boxes). **b,** A comparision of predicted and actual ages for all individuals based on the full ageing model. **c,** Out-of-sample predictions for individuals in the validation cohort. **d,** Apparent methylomic ageing rate (AMAR) for each individual, based on the full ageing model without clinical variables. The distribution of ageing rates shows faster ageing for men than women.

To validate this model, we sought and obtained an additional 180 independent samples consisting of 39 Caucasian and 141 Hispanic individuals. These samples were processed in the same manner as the original samples, then used to predict age based on the original model (i.e., as trained on the original cohort). The predictions were highly accurate, with a correlation between age and predicted age of 91% and an error of 4.6 years (Figure 3.4c). The validation cohort also supported our findings for the two signatures of ageing. Of the markers with an age-associated change in methylation fraction in the primary cohort, 20,005 (31%) were supported in the validation cohort ($P < 0.05$, Supplementary Figure 2). The new data also reproduced the trend for increasing deviance (Binomial $P \sim 0$, Methods), the trend towards more moderate methylation fractions with age (Binomial $P < 10^{-116}$), and the corresponding increase in methylome entropy ($P = 5x10^{-3}$). Finally, the validation cohort reproduced our earlier finding that CpG islands had increased methylation with age (Kruskal-Wallis test $P < 10^{-226}$) with only half the deviance of other regions, and increased deviance at the shores.

Chapter 3.3.5 Methylome ageing rate associations

While the ageing model is able to predict age with high accuracy, it is perhaps just as valuable as a tool for identifying individuals who do not follow the expectation. For example, Figure 3.4b highlights two individuals which appeared to age at different rates than the rest of the population. These deviations are possibly a combination of statistical or measurement error and biological differences which reflect diversity in

methylome ageing rates. To examine this, we sought to use the ageing model to quantify each individual's *apparent methylomic ageing rate* (AMAR), defined as the ratio of their predicted age, based on methylation data, to their chronological age. Using AMAR we tested whether the model could distinguish ageing rates for possibly relevant clinical factors. We found that gender and BMI had significant contributions to ageing rate (F-test, $P = 3 \times 10^{-9}$, $P = 2 \times 10^{-3}$, Methods). In our cohort, the methylome of men appear to age approximately 4% faster than women (Figure 3.4d) and each point of BMI increases the ageing rate by 0.2%. We note that the distribution of ages for men and women were very similar ($P > 0.1$, KS-test). However, as BMI has an established correlation with age, we tested for the contribution of age-adjusted BMI, which was not found to significantly influence ageing rate. Likewise, the validation cohort confirmed the increased ageing rate for men ($P < 0.01$), but was inconclusive for BMI ($P > 0.05$).

As genetic associations have been previously reported with human longevity and ageing phenotypes[108–111], we examined whether the model could distinguish ageing rates for individuals with different genetic variants. For this purpose, we obtained 15x whole-exome sequences for 252 of the individuals in our methylome study. After sequence processing and quality control, these sequences yielded 10,694 common single nucleotide variants across the population (Methods). As a negative control, we confirmed that none of these variants were significant predictors of age, which is to be expected since the genome sequence is considered to be relatively static over the course of a lifetime. On the other hand, one might expect to find genetic

variants that modulate the methylation of age-associated markers, i.e. methylation quantitative-trait loci or meQTLs[112]. Testing each genetic variant for association with the top age-associated methylation markers, we identified 248 meQTLs (Methods, FDR < 0.05, Figure 3.5a). For validation, we selected 15 genetic variants—corresponding to 37 meQTLs— to test in an additional 325 individuals. Analysis of this validation cohort found seven genetic variants in seven corresponding meQTLs to be significant in the validation cohort (FDR < 0.05, Supplementary Table 5). While some of these SNPs acted in a *cis* relationship, we confirmed that none of these methylation markers had observed genetic variants which directly modified the CpG site.

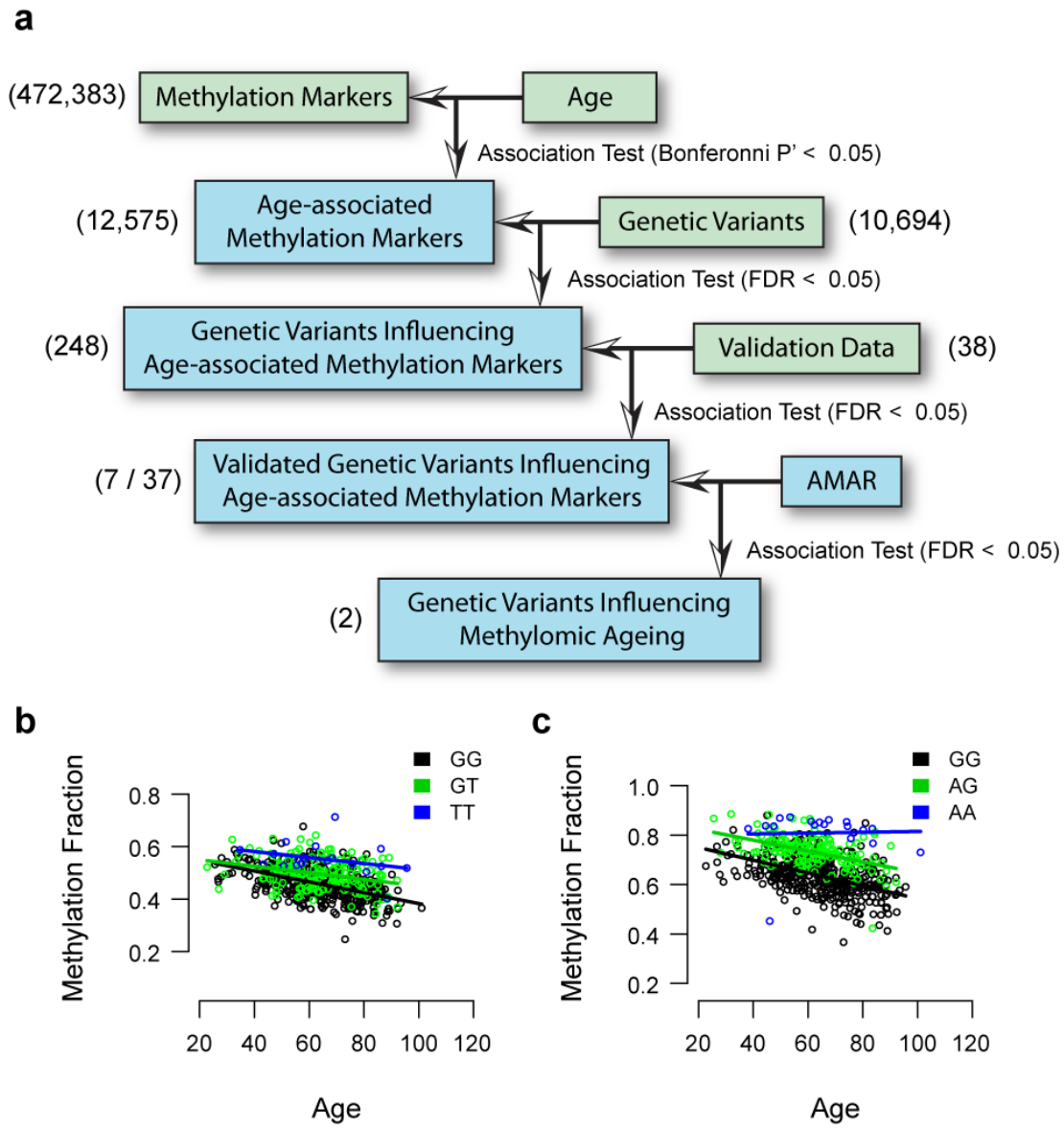


Figure 3.5 Genetic effects on methylomic ageing

a, We surveyed genomic variants for an association with age-associated methylation markers. 15 genetic variants, corresponding to 37 meQTLs, were chosen for validation. Of these, 7 were significant in the validation cohort and two showed an association with AMAR. **b,** A plot of the trend between the methylation marker cg27367526 (STEAP2) and age. The state of variant rs42663 (GTPBP10) causes an offset in this relationship. **c,** A second example for cg07906193 and rs17152433 (CTBP2, ZRANB1).

The methylation marker cg27193080 was one of those found to be significantly associated with age ($P < 10^{-19}$), and its methylation fraction was found to

be influenced by the SNP rs140692 ($P < 10^{-18}$) (Figure 3.5b). This meQTL was particularly interesting as both the SNP and the methylation marker mapped to the gene *methyl-CpG binding domain protein 4* (MBD4), one of the few known genes encoding a protein that can bind to methylated DNA. This meQTL thus captures a cis-relationship in which rs140692 influences the methylation state of MBD4. That MBD4 plays a role in human ageing is supported by previous work linking MBD4 to DNA repair, as well as work showing that mutations and knock-downs of MBD4 lead to increased genomic instability[113,114].

Of these validated meQTLs, two were identified that had a statistically-significant association with ageing rate (AMAR, FDR < 0.05, Figure 3.5b,c). One is the genetic marker rs17152433, which has a cis effect on the methylation marker cg07906193 near the genes CTBP2 and ZRANB1. Variants in this region have been previously shown to associate with increased incidence of prostate cancer[115]. The second genotype found to influence AMAR was rs42663 in the gene GTPBP10, which was associated with cg27367526 in the gene STEAP2. STEAP2 is known to play a role in maintenance of iron and copper homeostasis— metals which serve as essential components of the mitochondrial respiratory chain[116]. Studies have shown that perturbations of iron concentrations can induce DNA damage through oxidative stress in mammalian cells[117,118]. These meQTLs represent genetic variants that appear to broadly influence the ageing methylome and may be good candidates for further age-associated disease and longevity research.

Chapter 3.4 Conclusions

In this study we have shown that genome-wide methylation patterns represent a remarkably strong and reproducible biomarker of ageing. These patterns enable a quantitative model of the ageing methylome which demonstrates high accuracy and an ability to discriminate relevant factors in ageing, including gender and genetic variants. The ability to accurately measure age from molecular biomarkers has many potential practical implications, from health assessment and prevention of disease to forensic analysis. Similar to the effect of gender in this study, the identification of additional biometric or environmental factors that influence AMAR, such as smoking, alcohol consumption, or diet, will permit quantitative assessments of their impacts on health and longevity. A useful example would be to periodically assess the rate of ageing of an individual using AMAR and determine if diet or environmental factors can accelerate or retard the ageing process. As models of human ageing improve, it is conceivable that biological age, as measured from molecular profiles, might one day supersede chronological age in the clinical evaluation and treatment of patients.

Chapter 3.5 Methods

Chapter 3.5.1 Sample collection and test procedures

This study was approved by the institutional review boards of the University of California, San Diego and the University of Southern California. All participants signed informed consent statements prior to participation. Blood was drawn from a vein in the patient's arm into blood collection tubes containing the anticoagulant acid

citrate dextrose. Genomic DNA was extracted from the whole blood using a Qiagen FlexiGene DNA Kit and stored at -20 degrees Celsius. Raw methylation values for the autosomal chromosomes were obtained using the Illumina Infinium HumanMethylation450 BeadChip[90] assay. Markers with a detection p-value greater than 0.01 were set to missing. Two samples and 1481 markers were removed as they had greater than 5% missing values. The remaining missing values were imputed with the KNN approach (10 nearest markers) using the R "impute" package[119]. We performed exome sequencing on 258 of these samples, using a solution hybrid selection method to capture DNA followed by parallel sequencing on an Illumina HiSeq platform. Genotype calls were made using the SOAP program[120]. Calls with a quality score less than twenty were set as missing. Only variants which had fewer than 10% missing calls, were within Hardy-Weinberg equilibrium ($P <= 10^{-4}$), and of a common frequency (> 5%) were retained (10,694). Individuals with less than 20% missing calls (252) were retained. Additional genotyping was done with multiplex PCR followed by MALDI-TOF mass spectrometry analysis using the iPLEX/MassARRAY/Typer platform.

Chapter 3.5.2 Methylation quality control

We used principal component (PC) analysis to identify and remove outlier samples. We converted each sample into a z-score statistic, based on the squared distance of its 1st PC from the population mean. The z-statistic was converted to a false-discovery rate using the Gaussian cumulative distribution and the Benjamini-

Hochberg procedure[121]. Samples falling below an FDR of 0.2 were designated at outliers and removed. This filtering procedure was performed iteratively until no samples were determined to be an outlier. A total of 10 samples were removed in this manner.

Chapter 3.5.3 Computing methylation deviance

Methylation deviance was computed using the following approach: First, we removed the methylation trends due to all given variables, including age, gender, and BMI by fitting a linear model for each marker and acting only on the residuals. Next, we identified and removed highly non-normal markers based on the Shapiro-Wilk test ($P < 10^{-5}$). To allow for naturally occurring extreme deviations in the normality test, we first estimated the outliers of each marker based on the FDR of a z-score statistic. If any samples had an FDR less than 0.4, we ignored them and repeated the outlier detection until no outliers were detected. Finally, the deviance of each remaining marker was computed as the square of its adjusted methylation value.

Chapter 3.5.4 Association testing

Association tests were performed using nested linear models and the F-test. As methylation levels may be sensitive to a number of factors, we included several covariates, including gender, BMI, diabetes status, ethnicity, batch, P05 red, and P05 green. P05 red and green represent the number of detected methylation values found in the red and green channels for each individual at a p-value cutoff of 0.05. Tests for

whole-methylome changes in deviance were computed using the binomial test, based on the number of markers with a positive rather than negative coefficient.

Chapter 3.5.5 Annotation enrichment

Methylation marker annotations for CpG islands and GO terms were obtained from the IlluminaHumanMethylation450k.db database from Bioconductor[122]. Annotation enrichment tests were performed using the two-sided Fisher's exact test.

Chapter 3.5.6 Entropy analysis

Entropy statistics were computed on methylation data adjusted for covariates and filtered for normality (see Computing Methylation Deviance). We computed the normalized Shannon entropy[95] of an individual's methylome according to the formula:

$$Entropy = \frac{1}{N * \log\left(\frac{1}{2}\right)} \sum_{i} [MFi * \log(MFi) + (1 - MFi) * \log\left(1 - MFi\right)]$$

where MFi is the methylation fraction of the i^{th} methylation marker and N is the number of markers.

Chapter 3.5.7 Mapping CpG islands

Genomic positions and marker annotations for 27,176 CpG islands were obtained from the IlluminaHumanMethylation450k.db database from Bioconductor[122]. We obtained the positions for markers within each island with at least four markers

(25,028), as well as the nearest 100 markers upstream and downstream. These positions were then combined with the marker value of interest (i.e. methylation fraction, ageing coefficient, deviance) to produce a genomic map for each island and the surrounding region. After normalizing each map to the center of the island, we averaged the values at each relative genomic point across all islands to produce a common map.

Chapter 3.5.8 Ageing model

The diagnostic model of age was made using a multivariate linear model approach based on the elastic net algorithm implemented in the R package 'glmnet'[123]. Optimal regularization parameters were estimated using cross-validation. Using bootstrap analysis (N = 500), we included only markers in the final model that were present in more than half of all bootstraps. Covariates were included in the model and were exempted from penalization (regularization). P-values are based on a least-squares model built using the same terms and drop-one F-tests. We considered that the significance of gender and BMI might be explained through simple age correlations. This was not the case for gender, as it was not associated with age (Kruskal-Wallis test $P > 0.5$). BMI was found to lose significance in the model when first adjusted to account for age. AMAR was computed using the ageing model, but without the variables gender, BMI, and diabetes status. The coefficients were not changed. AMAR was then taken as an individual's predicted age divided by her or his actual age.

Chapter 3.5.9 Genetic variant associations

Each genetic variant was tested for association in an additive model with the top ageing associated methylation markers using nested linear models and the F-test. The same covariates used for age associations were included. Variant positions were based on the human reference build GRCh37 and gene annotations were based on chromosomal proximity within 20kbp.

Chapter 3.6 Methods summary

Methylation values were obtained using the Illumina Infinium HumanMethylation450 BeadChip[90] assay[119]. Exome sequencing was performed by parallel sequencing on an Illumina HiSeq platform. Genotype calls were made using the SOAP program[120]. Association tests were performed using nested linear models and the F-test. We included the covariates gender, BMI, diabetes status, ethnicity, batch, source location, and assay summary statistics. Multiple-hypothesis corrections were performed using the Benjamini Hochberg procedure[121]. Methylation deviance was computed in three steps: First, we removed the trends due to all given variables, including age, gender, BMI, and batch by fitting a linear model for each marker and acting only on the residuals. Next, we identified and removed highly non-normal markers. Finally, the deviance of each marker was computed as the square of its residual methylation value. Methylation marker annotations for CpG islands and GO terms were obtained from the IlluminaHumanMethylation450k.db database[122]. Annotation enrichment tests were performed using the two-sided Fisher's exact test.

Gene annotations were based on chromosomal proximity within 20kbp. Genomic positions for methylation markers were combined with marker values of interest (i.e. methylation fraction, ageing coefficient, deviance) to produce a genomic map for each CpG island. After normalizing each map to the center of the island, we averaged the values across all islands to produce a common map. Shannon entropy[95] statistics were computed on methylation data adjusted for covariates and filtered for normality. The diagnostic model of age was built using the elastic net algorithm[123]. Regularization parameters were estimated using cross-validation and bootstrap analysis. Covariates were included in the model and were exempted from penalization. P-values are based on a least-squares model built using the same terms.

Chapter 3.7 Supplementary information

Supplementary information is linked to the online version of the paper at www.nature.com/nature.

Chapter 3.8 Acknowledgments

Hannum G, Guinney J, Zhang L, Zhao L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Rajapakse I, Friend S, Ideker T, Zhang K. (2012) Genome-wide Methylation Profiles Reveal Quantitative Views of Human Ageing Rates. (Presently in submission at Nature)

We thank Janusz Dutkowski and Mariano Alvarez for critical discussions and Daniel O'Conner for reviewing the manuscript. G.Ha and T.I are supported by NIH

grants P50GM085764 and R01E5014811. JG is supported by the NIH grants 3104672 and U54CA149237. K. Z. G.Hu., L.Zhao, and L.Zhang were supported by grants from Chinese National 985 Project to Sichuan University and West China Hospital, NEI/NIH grants EY014428, EY018660, EY019270, EY021374, the VA Merit Award, the Research to Prevent Blindness, and the Burroughs Wellcome Fund Clinical Scientist Award in Translational Research. This work is a product of the Sage Federation, a consortium of research labs whose goal is to encourage greater openness and sharing of biomedical data and analyses.

Gregory and Justin contributed equally to this work. Author contributions: G.Hu., L.Zhao, L.Zhang, S.S, and Y.G. collected and processed samples with guidance from K.Z. B.K, M.B, and J.F. performed the methylation assays. G.Ha. and J.G. performed the principal statistical analyses with guidance from T.I., R.D., and S.F. I.R. discussed the entropy metric. G.Ha., J.G., T.I., Y.G., and K.Z. wrote the manuscript.

 B.K, M.B, and J.F. work for Illumina Inc. Correspondence and requests for information should be addressed to T.I. (tideker@ucsd.edu) or K.Z. (kzhang@ucsd.edu).

BIBLIOGRAPHY

1. Wilmes, G.M. *et al.* A genetic interaction map of RNA-processing factors reveals links between Sem1/Dss1-containing complexes and mRNA export and splicing. *Molecular Cell* **32**, 735-746 (2008).

2. Collins, S.R. *et al.* Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806-810 (2007).

3. Pan, X. *et al.* A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell* **124**, 1069-1081 (2006).

4. Tong, A.H.Y. *et al.* Global mapping of the yeast genetic interaction network. *Science (New York, N.Y.)* **303**, 808-813 (2004).

5. Brem, R.B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 1572-1577 (2005).

6. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nature Genetics* **39**, 1217-1224 (2007).

7. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

8. Boone, C., Bussey, H. & Andrews, B.J. Exploring genetic interactions and networks with yeast. *Nat Rev Genet* **8**, 437-449 (2007).

9. Collins, S.R., Schuldiner, M., Krogan, N.J. & Weissman, J.S. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biology* **7**, R63 (2006).

10. Roguev, A. *et al.* Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science (New York, N.Y.)* **322**, 405-410 (2008).

11. Schuldiner, M. *et al.* Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507-519 (2005).

12. Primrose, S. & Twyman, R. *Principles of gene manipulation and genomics.* xxii, 644 (Oxford: Blackwell Publishing: Malden, MA, 2006).

13. Huang, Y. *et al.* Exploiting gene x gene interaction in linkage analysis. *BMC Proceedings* **1 Suppl 1**, S64 (2007).

14. Evans, D.M., Marchini, J., Morris, A.P. & Cardon, L.R. Two-stage two-locus models in genome-wide association. *PLoS Genetics* **2**, e157 (2006).

15. Marchini, J., Donnelly, P. & Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**, 413-417 (2005).

16. Litvin, O., Causton, H.C., Chen, B.-J. & Pe'er, D. Modularity and interactions in the genetics of gene expression. *Proceedings of the National Academy of Sciences* **106**, 6441-6446 (2009).

17. Storey, J.D., Akey, J.M. & Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biology* **3**, e267 (2005).

18. Wall, J.D. & Pritchard, J.K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews. Genetics* **4**, 587-597 (2003).

19. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology* **23**, 561-566 (2005).

20. Gavin, A.-C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631-636 (2006).

21. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Research* **30**, 31-34 (2002).

22. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25-29 (2000).

23. Storey, J.D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 479-498 (2002).

24. Christman, M.F., Dietrich, F.S. & Fink, G.R. Mitotic recombination in the rDNA of S. cerevisiae is suppressed by the combined action of DNA topoisomerases I and II. *Cell* **55**, 413-425 (1988).

25. Kressler, D., Linder, P. & de La Cruz, J. Protein trans-acting factors involved in ribosome biogenesis in Saccharomyces cerevisiae. *Molecular and Cellular Biology* **19**, 7897-7912 (1999).

26. Koehler, C.M. *et al.* Tim9p, an essential partner subunit of Tim10p for the import of mitochondrial carrier proteins. *The EMBO Journal* **17**, 6477-6486 (1998).

27. Jungmann, J. & Munro, S. Multi-protein complexes in the cis Golgi of Saccharomyces cerevisiae with alpha-1,6-mannosyltransferase activity. *The EMBO Journal* **17**, 423-434 (1998).

28. Sacher, M., Barrowman, J., Schieltz, D., Yates, J.R. 3rd & Ferro-Novick, S. Identification and characterization of five new subunits of TRAPP. *European Journal of Cell Biology* **79**, 71-80 (2000).

29. Iung, A.R. *et al.* Mitochondrial function in cell wall glycoprotein synthesis in Saccharomyces cerevisiae NCYC 625 (Wild type) and [rho(0)] mutants. *Applied and Environmental Microbiology* **65**, 5398-5402 (1999).

30. Ulitsky, I. & Shamir, R. Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. *Molecular Systems Biology* **3**, 104 (2007).

31. Shimada, K. *et al.* Ino80 chromatin remodeling complex promotes recovery of stalled replication forks. *Current Biology: CB* **18**, 566-575 (2008).

32. van Attikum, H., Fritsch, O., Hohn, B. & Gasser, S.M. Recruitment of the INO80 complex by H2A phosphorylation links ATP-dependent chromatin remodeling with DNA double-strand break repair. *Cell* **119**, 777-788 (2004).

33. Shen, X., Mizuguchi, G., Hamiche, A. & Wu, C. A chromatin remodelling complex involved in transcription and DNA processing. *Nature* **406**, 541-544 (2000).

34. Papamichos-Chronakis, M. & Peterson, C.L. The Ino80 chromatin-remodeling enzyme regulates replisome function and stability. *Nature Structural & Molecular Biology* **15**, 338-345 (2008).

35. Schwabish, M.A. & Struhl, K. Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. *Molecular and Cellular Biology* **24**, 10111-10117 (2004).

36. Ford, J., Odeyale, O., Eskandar, A., Kouba, N. & Shen, C.-H. A SWI/SNF- and INO80-dependent nucleosome movement at the INO1 promoter. *Biochemical and Biophysical Research Communications* **361**, 974-979 (2007).

37. Klopf, E. *et al.* Cooperation between the INO80 complex and histone chaperones determines adaptation of stress gene transcription in the yeast S. cerevisiae. *Mol. Cell. Biol.* MCB.01858-08 (2009).doi:10.1128/MCB.01858-08

38. Carlborg, O. & Haley, C.S. Epistasis: too often neglected in complex trait studies? *Nature Reviews. Genetics* **5**, 618-625 (2004).

39. Jonikas, M.C. *et al.* Comprehensive Characterization of Genes Required for Protein Folding in the Endoplasmic Reticulum. *Science* **323**, 1693-1697 (2009).

40. Warner, J.R. Synthesis of ribosomes in Saccharomyces cerevisiae. *Microbiological Reviews* **53**, 256-271 (1989).

41. Brauer, M.J. *et al.* Coordination of Growth Rate, Cell Cycle, Stress Response, and Metabolic Activity in Yeast. *Mol. Biol. Cell* **19**, 352-367 (2008).

42. Schadt, E.E. & Lum, P.Y. Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *Journal of Lipid Research* **47**, 2601-2613 (2006).

43. Rockman, M.V. & Kruglyak, L. Genetics of global gene expression. *Nature Reviews. Genetics* **7**, 862-872 (2006).

44. Suthram, S., Beyer, A., Karp, R.M., Eldar, Y. & Ideker, T. eQED: an efficient method for interpreting eQTL associations using protein networks. *Molecular Systems Biology* **4**, 162 (2008).

45. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nature Reviews. Genetics* **10**, 241-251 (2009).

46. Cherry, J.M. *et al.* Genetic and physical maps of Saccharomyces cerevisiae. *Nature* **387**, 67-73 (1997).

47. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**, 2498-2504 (2003).

48. Cohen, B.A., Mitra, R.D., Hughes, J.D. & Church, G.M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genetics* **26**, 183-186 (2000).

49. Schuldiner, M., Collins, S.R., Weissman, J.S. & Krogan, N.J. Quantitative genetic analysis in Saccharomyces cerevisiae using epistatic miniarray profiles

(E-MAPs) and its application to chromatin functions. *Methods (San Diego, Calif.)* **40**, 344-352 (2006).

50. Beyer, A., Bandyopadhyay, S. & Ideker, T. Integrating physical and genetic maps: from genomes to interaction networks. *Nature Reviews. Genetics* **8**, 699-710 (2007).

51. Fiedler, D. *et al.* Functional organization of the S. cerevisiae phosphorylation network. *Cell* **136**, 952-963 (2009).

52. Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to DNA damage. *Science (New York, N.Y.)* **330**, 1385-1389 (2010).

53. Costanzo, M. *et al.* The Genetic Landscape of a Cell. *Science* **327**, 425-431 (2010).

54. Schlabach, M.R. *et al.* Cancer proliferation gene discovery through functional genomics. *Science (New York, N.Y.)* **319**, 620-624 (2008).

55. Bakal, C. *et al.* Phosphorylation networks regulating JNK activity in diverse genetic backgrounds. *Science (New York, N.Y.)* **322**, 453-456 (2008).

56. Breitkreutz, B.-J. *et al.* The BioGRID Interaction Database: 2008 update. *Nucleic Acids Research* **36**, D637-640 (2008).

57. Zhang, L.V. *et al.* Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. *Journal of Biology* **4**, 6 (2005).

58. Bandyopadhyay, S., Kelley, R., Krogan, N.J. & Ideker, T. Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Computational Biology* **4**, e1000065 (2008).

59. Jaimovich, A., Rinott, R., Schuldiner, M., Margalit, H. & Friedman, N. Modularity and directionality in genetic interaction maps. *Bioinformatics* **26**, i228 -i236 (2010).

60. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols* **2**, 2366-2382 (2007).

61. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)* **27**, 431-432 (2011).

62. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)* **21**, 3448-3449 (2005).

63. Ashkenazi, M., Bader, G.D., Kuchinsky, A., Moshelion, M. & States, D.J. Cytoscape ESP: simple search of complex biological networks. *Bioinformatics (Oxford, England)* **24**, 1465-1466 (2008).

64. van Iersel, M.P. *et al.* The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* **11**, 5 (2010).

65. Saeed, A.I. *et al.* TM4 microarray software suite. *Methods in Enzymology* **411**, 134-193 (2006).

66. Collins, S.R. *et al.* Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & Cellular Proteomics: MCP* **6**, 439-450 (2007).

67. Pu, S., Vlasblom, J., Emili, A., Greenblatt, J. & Wodak, S.J. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics* **7**, 944-960 (2007).

68. Pu, S., Wong, J., Turner, B., Cho, E. & Wodak, S.J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research* **37**, 825-831 (2009).

69. Hang, M. & Smith, M.M. Genetic analysis implicates the Set3/Hos2 histone deacetylase in the deposition and remodeling of nucleosomes containing H2A.Z. *Genetics* **187**, 1053-1066 (2011).

70. Steven N., A. Why women live longer than men: Sex differences in longevity. *Gender Medicine* **3**, 79-92 (2006).

71. Blair, S.N. *et al.* Physical fitness and all-cause mortality. A prospective study of healthy men and women. *JAMA: The Journal of the American Medical Association* **262**, 2395-2401 (1989).

72. Harley, C.B., Futcher, A.B. & Greider, C.W. Telomeres shorten during ageing of human fibroblasts. *Nature* **345**, 458-460 (1990).

73. Epel, E.S. *et al.* Accelerated telomere shortening in response to life stress. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 17312 -17315 (2004).

74. Valdes, A. *et al.* Obesity, cigarette smoking, and telomere length in women. *The Lancet* **366**, 662-664

75. Fraser, H.B., Khaitovich, P., Plotkin, J.B., Pääbo, S. & Eisen, M.B. Aging and Gene Expression in the Primate Brain. *PLoS Biol* **3**, e274 (2005).

76. Zahn, J.M. *et al.* AGEMAP: A Gene Expression Database for Aging in Mice. *PLoS Genet* **3**, e201 (2007).

77. de Magalhães, J.P., Curado, J. & Church, G.M. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* **25**, 875 -881 (2009).

78. Fraga, M.F. & Esteller, M. Epigenetics and aging: the targets and the marks. *Trends in Genetics: TIG* **23**, 413-418 (2007).

79. Christensen, B.C. *et al.* Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *PLoS Genet* **5**, e1000602 (2009).

80. Bollati, V. *et al.* Decline in Genomic DNA Methylation through Aging in a Cohort of Elderly Subjects. *Mechanisms of ageing and development* **130**, 234-239 (2009).

81. Boks, M.P. *et al.* The Relationship of DNA Methylation with Age, Gender and Genotype in Twins and Healthy Controls. *PLoS ONE* **4**, e6767 (2009).

82. Rakyan, V.K. *et al.* Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Research* (2010).doi:10.1101/gr.103101.109

83. Bocklandt, S. *et al.* Epigenetic Predictor of Age. *PLoS ONE* **6**, e14821 (2011).

84. Barres, R. & Zierath, J.R. DNA methylation in metabolic disorders. *The American Journal of Clinical Nutrition* (2011).doi:10.3945/ajcn.110.001933

85. Jones, P.A. & Laird, P.W. Cancer epigenetics comes of age. *Nature Genetics* **21**, 163-167 (1999).

86. Esteller, M. Epigenetics in Cancer. *The New England Journal of Medicine* **358**, 1148-1159 (2008).

87. Fraga, M.F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10604 -10609 (2005).

88. Vijg, J. & Campisi, J. Puzzles, promises and a cure for ageing. *Nature* **454**, 1065-1071 (2008).

89. Murgatroyd, C. *et al.* Dynamic DNA methylation programs persistent adverse effects of early-life stress. *Nature Neuroscience* **12**, 1559-1566 (2009).

90. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-295 (2011).

91. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).

92. Ng, C.-F. *et al.* FHL2 exhibits anti-proliferative and anti-apoptotic activities in liver cancer cells. *Cancer Letters* **304**, 97-106 (2011).

93. Fan, M. *et al.* Diverse Gene Expression and DNA Methylation Profiles Correlate with Differential Adaptation of Breast Cancer Cells to the Antiestrogens Tamoxifen and Fulvestrant. *Cancer Research* **66**, 11954 -11966 (2006).

94. Breusch, T.S. & Pagan, A.R. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* **47**, 1287 (1979).

95. Shannon, C.E. & Weaver, W. *The Mathematical Theory of Communication*. *The Mathematical Theory of Communication* (University of Illinois Press: Champaign, IL., 1963).

96. Sun, Z. *et al.* Integrated Analysis of Gene Expression, CpG Island Methylation, and Gene Copy Number in Breast Cancer Cells by Deep Sequencing. (2011).doi:10.1371/journal.pone.0017490

97. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423-428 (2008).

98. Southworth, L.K., Owen, A.B. & Kim, S.K. Aging mice show a decreasing correlation of gene expression within genetic modules. *PLoS genetics* **5**, e1000776 (2009).

99. Bahar, R. *et al.* Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature* **441**, 1011-4 (2006).

100. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-775 (2011).

101. Irizarry, R.A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics* **41**, 178-186 (2009).

102. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301-320 (2005).

103. Yacubova, E. & Komuro, H. Stage-specific control of neuronal migration by somatostatin. *Nature* **415**, 77-81 (2002).

104. Saito, T. *et al.* Somatostatin regulates brain amyloid beta peptide Abeta42 through modulation of proteolytic degradation. *Nature Medicine* **11**, 434-439 (2005).

105. Small, K.S. *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature genetics* **43**, 561-4 (2011).

106. Lane, M.A. *et al.* Calorie restriction lowers body temperature in rhesus monkeys, consistent with a postulated anti-aging mechanism in rodents. *Proceedings of the National Academy of Sciences* **93**, 4159 -4164 (1996).

107. Tatar, M., Bartke, A. & Antebi, A. The endocrine regulation of aging by insulin-like signals. *Science (New York, N.Y.)* **299**, 1346-51 (2003).

108. Atzmon, G. *et al.* Lipoprotein Genotype and Conserved Pathway for Exceptional Longevity in Humans. *PLoS Biol* **4**, e113 (2006).

109. Suh, Y. *et al.* Functionally significant insulin-like growth factor I receptor mutations in centenarians. *Proceedings of the National Academy of Sciences* **105**, 3438 -3442 (2008).

110. Willcox, B.J. *et al.* FOXO3A genotype is strongly associated with human longevity. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 13987-13992 (2008).

111. Wheeler, H.E. *et al.* Sequential Use of Transcriptional Profiling, Expression Quantitative Trait Mapping, and Gene Association Implicates MMP20 in Human Kidney Aging. *PLoS Genet* **5**, e1000685 (2009).

112. Bell, J.T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology* **12**, R10 (2011).

113. Bellacosa, A. *et al.* MED1, a novel human methyl-CpG-binding endonuclease, interacts with DNA mismatch repair protein MLH1. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 3969-74 (1999).

114. Bertoni, C., Rustagi, A. & Rando, T.A. Enhanced gene repair mediated by methyl-CpG-modified single-stranded oligonucleotides. *Nucleic acids research* **37**, 7468-82 (2009).

115. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics* **40**, 310-315 (2008).

116. Ohgami, R.S., Campagna, D.R., McDonald, A. & Fleming, M.D. The Steap proteins are metalloreductases. *Blood* **108**, 1388-94 (2006).

117. Hartwig, A. & Schlepegrell, R. Induction of oxidative DNA damage by ferric iron in mammalian cells. *Carcinogenesis* **16**, 3009-3013 (1995).

118. Karthikeyan, G., Lewis, L.K. & Resnick, M.A. The mitochondrial protein frataxin prevents nuclear damage. *Human molecular genetics* **11**, 1351-62 (2002).

119. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics (Oxford, England)* **17**, 520-525 (2001).

120. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)* **24**, 713-714 (2008).

121. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).

122. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80 (2004).

123. Jerome Friedman, Trevor Hastie, R.T. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, (2010).