It Takes Two to Tango: Genotyping and Phenotyping in Genome-Wide Association Studies

Ohad Nachtomy
Department of Philosophy
Bar-Ilan University
Ramat-Gan, Israel
&
Department of Philosophy
Princeton University
Princeton, NJ, USA
ohadnachtomy@mac.com

Yaron Ramati
Department of Humanities and Arts
Technion–Israel Institute of Technology
Haifa, Israel
huramati@technion.ac.il

Ayelet Shavit
Tel-Hai Academic College
Upper Galilee, Israel

Zohar Yakhini
Department of Computer Science
Technion–Israel Institute of Technology
Haifa, Israel

Abstract
In this article we examine the "phenotype" concept in light of recent technological advances in Genome-Wide Association Studies (GWAS). By observing the technology and its presuppositions, we put forward the thesis that at least in this case genotype and phenotype are effectively coidentified one by means of the other. We suggest that the coidentification of genotype–phenotype couples in expression-based GWAS also indicates a conceptual dependence, which we call "co-definition." We note that viewing these terms as codefined runs against possible expectations, viz., that genotypes and phenotypes could ultimately be expressed independently from one another. In addition, the co-definition of genotypes and phenotypes in this context emphasizes the correlative (rather than mechanistic) character of both genotypes and phenotypes in GWAS.

Keywords
genetics, genome-wide association, genotype, Human Genome Project, phenotype, regulatory genes

August 20, 2009; revised and accepted October 31, 2009
Biological Theory 4(3) 2009, 294–301.

In a recent article (Nachtomy et al. 2007), we have shown that the increasing power of current microarray measurement technologies results in an extension of the concept of the phenotype. We stress that this is not a revolution, but a significant extension of the molecular dimension of the concept of the phenotype. In particular, the extended concept is more comprehensively quantitative and grounded in complex molecular properties. Thus, for example, in Genome-Wide Association Studies (GWAS), the abundance of an RNA transcript is regarded as a phenotypic trait.

In the present study we look more closely at the way in which such phenotypes are related to genotypes in GWAS and observe some conceptual consequences of this relation. In closely examining the way in which the practitioners use and advance the technology of phenotyping and genotyping, we conclude that some of the most advanced technologies in current molecular biology refer us back to the formative conceptualizations of genotype and phenotype relations early in the 20th century.

Historically, the existence of regulatory DNA sequences such as transcription factor binding sites that affect RNA expression was realized almost in parallel to the appearance of the idea of the molecular gene (Keller 2002). The *lac* operon model, explored in the 1960s by Jacques Monod and François Jacob (Jacob and Monod 1961), suggested a mechanistic model by which outside stimuli controlled the abundance of the mRNA transcript for a specific protein. A DNA region next to the coding site for β-galactosidase was found to be one part of a regulatory mechanism for the transcription of the enzyme's mRNA, dependent on the presence of glucose or lactose in the immediate vicinity of *E. coli*. Rather than looking for regulatory proteins' binding sites on the DNA of the form suggested in the operon model, expression-based GWAS look for differentially regulated phenotypes. Instead of looking for mechanisms of operation, expression-based GWAS search for loci in the DNA that are functionally significant for RNA abundance, significantly challenging the dominant mechanistic conceptualization of molecular genetics.

For T. H. Morgan, the hereditary factor was not the bearer of traits, as suggested before him by William Bateson, but rather a factor that produces changes in the observed phenotype.[1] Morgan suggested viewing the gene not as determining the eye color of Drosophila, but rather viewing the different alleles as affecting the color of the eye. Morgan's conception of what was later termed the *functional gene* was revoked only by the discovery of DNA and the rise of molecular genetics (Weber 2005).

The classical molecular approach suggested a molecular structure for the gene, a mechanism of RNA transcription and protein translation, as well as regulation of RNA transcription. In Waters' (2000) words, the initial identification of the molecular gene was but "the image in the DNA" of the molecule (RNA or polypeptide), whose biological activity is of interest to the experimenter, essentially combining the competing structural and functional conceptualizations of the gene that were prevalent before 1953. Notwithstanding, the "consensus gene" concept in contemporary genetics, though defined and characterized by reference to functional sequence elements, is "identified strictly from physical readouts of the DNA sequences" (Fogle 2001: 3–25).[2] As Lewontin (1992: 143) points out in reference to developments in molecular genetics, "the developments of techniques of observing the phenotype have been revolutionary for genetic analysis, precisely because they solve the problem of inferring genotype from phenotype by eliminating development. All genotypes, irrespective of their influence on development, can be unambiguously discriminated at the molecular level" of the phenotype. The identification of gene structures in the DNA is thus effectively made independent of any phenotypic effect. The expression-based GWAS method of investigation provides a clear example of an alternative process of identifying DNA sequence hereditary units in the molecular era that goes beyond the classical concept of the molecular gene. Unlike the RNA-coding areas on the DNA that are structurally identified through common features, the identification of the regulatory locus in GWAS studies is purely functional and makes no commitment to a structure: Like Morgan's functional *gene*, the genotype differentially regulating RNA transcript abundance is recognized through its effects on the phenotype independently of its base sequence.

We here demonstrate how the treatment of expression profiles as phenotypes actually makes reference to genotypes, and how these genotypes make reference to their presumed phenotypes. The link between genotype and phenotype within this method of investigation can be stated thus: In expression-based GWAS, *genotype and phenotype co-identify each other*.[3]

While this finding is similar to Waters' (2000) "image in the DNA" of the classical molecular gene, the coidentification of the genotype–phenotype couple (G/P couple) we describe here goes a step forward in suggesting that not only the genotype is identified through the phenotype but the phenotype is also simultaneously identified. Whereas Waters' classical gene has its phenotype given as the amino acid or nucleic acid linear sequence of the "molecule of interest," independent of any reference to the genotype, we here point to two conceptual diversions: (1) significant molecular phenotypes are quantitatively characterized; (2) the molecular phenotype is coidentified with the genotype. This, we believe, is a significant observation on the concept of the phenotype. Within expression-based GWAS, not only does the genotype hold a conceptual relation to the phenotype but the phenotype also holds a reciprocal conceptual relation to the genotype. In effect, the coidentification of the genotype and phenotype implies their co-definition by each other.

The scientific methodology we observe in GWAS is not marginal but constitutes one of the most rapidly advancing and central fronts of molecular biology. More than 100 GWAS were conducted in recent years on human cohorts, identifying hundreds of polymorphisms (Goldstein 2009), bringing about "one of the most prolific periods of discovery in human genetics" (Hirschhorn 2009: 1699–1701). The mutual dependency of genotype and phenotype observed in the context of this scientific methodology demonstrates that a leading method of investigation in contemporary molecular genetics regards the genotype as being codefined together with its corresponding phenotype.

A clarification of the scope of our claim is in order. We do not make a general claim for the mutual dependence between genotype and phenotype. Instead, we limit the scope of this study to the current practice of expression-based GWAS. Our methodology consists of studying the way these two concepts (genotype and phenotype) are exemplified and employed in expression-based GWAS. We find the possibility of extending this thesis to other branches of biology and medicine intriguing, but we do not pursue it here. We do think, however, that if genotype and phenotype are indeed mutually dependent in this central case of contemporary genetics, this indicates a shift in the conceptual perception of some scientists, and we believe that this possibility should be examined in other contexts as well.

In the first section, we present the scientific approach of expression-based GWAS and of the technologies it employs. We then present the conceptual underpinning of this scientific approach, viz., the coidentification of genotypes and phenotypes. In the final section, we discuss some implications of these observations.

Expression-Based Genome-Wide Association Studies

The abundance of an RNA transcript is a quantitative trait, which like any other biological trait, can be treated as a phenotype (Morley et al. 2004; Nachtomy et al. 2007). A revolutionary technology of gene expression profiling, introduced to biology in the last decade, enables researchers to measure RNA transcript abundance in different cell populations, including single cell organisms, tissue samples, and cell lines. Using this technology it is now possible to assess the abundance of many transcripts, indeed, of the entire known range of RNA molecules (the *transcriptome*), simultaneously in the same sample, allowing gene expression abundance to be studied on a large and unbiased set of traits, in what is currently known as *transcriptomics*.

The strong tendency of nucleic acids of complementary sequence to react with each other, allows for the usage of populations of small RNA molecules, each of which is complementary to part of one or more of the genes in the genome. These probe molecules can be spotted as an array onto a slide and are now provided with a fluorescent tag that lights up when an RNA molecule hybridizes. The RNA is extracted from each individual or strain in the test population, and the abundance of each transcript is typically measured by hybridization to microarrays. This nucleotide array technology allows determining the transcription abundance of all known RNA transcripts.

Reducing heterogeneity and possible environmental effects is called for, in order to foreground the hereditary nature of RNA transcript abundance. Narrowing heterogeneity in the tested population, be it cell type, tissue, or single cell organism, allows the screening off of possible differences in transcript abundance among different cell types. Such narrowing of heterogeneity is enhanced by screening off other sources of variance, such as age, sex, and specific physiological conditions of the individuals (e.g., Emilsson et al. 2008). By using more than one sample from the same individual in the tested population, environmental and other transitory effects are also screened off. Any RNA transcript, whose abundance shows higher variance among different samples from the same individual than its variance among individuals, is eliminated from further analysis (e.g., Morley et al. 2004), leaving those RNA transcripts whose abundance is relatively constant in the individual.

Microarrays were first applied to the study of genetically based variation of transcript abundance showing differences in levels of gene expression among different strains of yeast (Brem et al. 2002) and mice (Sandberg et al. 2000). Demonstrating that such differences segregate in crosses (the median proportion of the observed genetic variation in a cross between haploid segregates of different strains of *Saccharomyces cerevisiae* was estimated to be 84%; Brem et al. 2002), these studies supported the assertion that at least some of the differences found in gene expression levels have a genetic background. Subsequent studies documented abundant heritable variation in transcript abundance in more than a dozen species, including Drosophila, killifish, maize, rats (reviewed by Rockman and Kruglyak 2006), and humans (Morley et al. 2004). A large cohort study of Icelandic population blood and adipose (fat) tissue samples (Emilsson et al. 2008) showed significant inheritance for RNA transcript abundance. In the blood tissue, 8,047 out of 23,720 expression traits have shown statistically significant pattern of heredity, while in adipose tissue 11,251 out of 23,720 expression traits have shown statistically significant pattern of heredity, with an average of 30% of the observed variation explained by inheritance, supporting findings from earlier studies (e.g., Dixon et al. 2007; Göring et al. 2007). Significant and consistent differences in transcript abundance were detected between different strains of Drosophila (Osada et al. 2006), and between different human ethnic groups (Spielman et al. 2007). The key finding of all these studies is that thousands of transcript-level traits show a consistent

pattern of inheritance, suggesting significant genetic influence over variation in transcript abundance.

Identification of the specific genotypic determinants responsible for the variation in transcript abundance can be performed using genome-wide association. GWAS represent a relatively new approach commonly used to identify genetic variation that influences phenotypic traits across the entire human genome (e.g., blood pressure or weight; see, e.g., Emilsson et al. 2008), or the presence or absence of a disease or condition (Office of Population Genomics 2009). As a direct outcome of the Human Genome Project, GWAS involve scanning genetic markers—either single-nucleotide polymorphism (SNPs), gene copy number variance (CNVs), or microsatellites—across the complete sets of DNA, or genomes, to find DNA variations associated with a particular phenotype. Single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring in a single nucleotide—A, T, C, or G in the genome of different members of the same species. As of 2009, more than a million SNPs have been identified in the human genome. It is estimated that since SNPs occur every 100 to 300 bases along the ~4-billion-base human genome, a total of 4 million SNPs exist in the entire human genome, almost all have only two variants in the population (Roeder and Luca 2009). Each GWA study measures thousands of SNPs at the same time in order to find those locations that are statistically associated with an increased risk of developing a certain disease, or any other phenotypic trait. Recently, expression-based GWA studies of the transcriptome have started to look at other types of variations in the human genome, mainly variation in a gene's copy number (e.g., Stranger et al. 2007). Although these studies consist of some unique features, they are not manifestly different from those addressing single nucleotide polymorphism.

In the context of transcriptomics, GWAS methods are used to identify genetic determinants implicated in modifying RNA transcript abundance (i.e., expression-quantitative trait loci or eQTLs; Schadt et al. 2003). Treated as a quantitative phenotype, transcript abundance of each of thousands of RNA molecules in the studied pool goes through a series of statistical tests in order to identify possible association to DNA polymorphism, detected in the studied population using SNP genetic libraries. These tests aim to identify possible eQTLs by revealing statistically significant correlation in the study population between the measured abundance of specific RNA transcripts tested and genetic polymorph markers. Association between a certain genomic locus and the abundance of a certain RNA transcript is established if it is shown that such correlation is found.

Indeed, published results of these association analysis studies demonstrate strong correlation between variations in transcript abundance and variation in DNA sequences (see Rockman and Kruglyak 2006 for review). In humans, these studies have shown that polymorphism in the DNA at the vicinity of the expressed genes (named *cis* regions) has a strong effect on the variation in the RNA abundance. For example, Emilsson and his colleagues (2008) showed that 2,529 and 1,307 (FDR = 0.05) out of 20,877 expression traits show at least one eQTL site in their vicinity in blood and adipose tissue, respectively. Similar findings were reported for maize (Schadt et al. 2003), rats (Hubner et al. 2005), yeast (Brem et al. 2002), and humans (Morley et al. 2004). These results suggest that variation in gene expression is due in large part to polymorphisms in the vicinity of the DNA coding region of genes. These studies indicate that genetically determined phenotypic diversity results not only from variation in DNA coding regions that affects the function of the coded proteins and functional RNA but also from regulatory variation that affects the abundance of the RNA transcript.

Recent studies suggest that macroscopic phenotypes are correlated to quantitative differences in RNA transcript abundance: changes in the expression levels of certain RNA transcripts were shown to be correlated with cancer (e.g., Hedenfalk et al. 2003), obesity (e.g., Emilsson et al. 2008), and other macroscopic phenotypic attributes. Moreover, genetic variation at the DNA-sequence level of individual members in a single species was shown to be directly implicated in observed variation in RNA transcript abundance. Taken together, these findings demonstrate more strongly than before the widespread prevalence of cases in which genetic variation affects the observed traits, not only through the sequence of the proteins, but also through its effect on the abundance of the different RNA transcripts.

We suggest that the conceptualization of the phenotype and genotype occurring in these correlation studies is of wider interest to the discussion on conceptualization in genetics. Through the case of expression-based GWAS, we argue for the *coidentification* of genotypes and phenotypes in contemporary molecular genetics.

Identifying and Differentiating Genotype and Phenotype in GWAS

An interesting conceptual consequence of the scientific method employed in GWA study of the transcriptome relates to the identification of genotypes and phenotypes and to the individuation of the latter. In the practical context of expression-based GWAS, *phenotypes and genotypes are coidentified*: A phenotype is identified as a correlate to polymorphic changes in the DNA; and a polymorphism in the DNA will be identified as a genotype in the functional sense only if it co-occurs with quantitative phenotypic differences. Furthermore, *the coidentified genotype further differentiates the quantitative phenotype*: The (typically) two genotypic polymorphs differentiate the coidentified phenotype into *high* and *low*

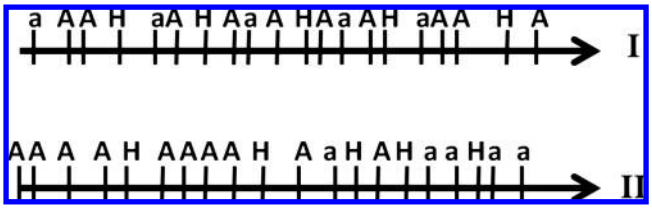

Figure 1.
An illustration of a superposition of results from 20 individuals of measured RNA abundance for one type of RNA strand (x-axis), and one SNP site (flag). A: Polymorph type I; a: polymorph type II; H: heterozygote. Case I illustrates lack of correlation between RNA abundance and the SNP. Case II illustrates possible correlation between DNA polymorphism type II and increase in RNA abundance. Expression-based GWAS test for statistical significant couplings in the form illustrated in case II within any pair of measured RNA abundance and SNP. Any significant correlation between RNA abundance and a DNA SNP site establishes a G/P couple.

subclasses, which become meaningful in the context of the GWA study. We will argue for these two observations in turn.

We would like to emphasize that the initial identification of the phenotypes is built into the technology of transcript abundance association studies. The microarray containing the different small RNA molecules *probes the transcriptome*, chemically testing it for the presence of the distinct RNA transcripts, so that each and every feature of the array selectively binds to a specific RNA transcript based on their nucleic base sequence. The output of this technology is a set of thousands of phenotypes; each feature of the microarray identifies a phenotype—an RNA transcript.

In this way, microarray technology does more than simply identify the presence of RNA molecules; for each of the probed molecules, this technology provides a continuous or semi-continuous range of quantitative values. This quantification of the RNA abundance constitutes a further division of the phenotype into a range of values, similar to the quantification of macro-phenotypic traits such as body height or weight. It is the quantitative variation in each of the phenotypes (rather than its presence/absence) that is the main goal of investigation for expression-based GWAS as discussed here.

A variety of technologies perform genome-wide determination of polymorphism. Most often, each testing kit provides probing of DNA polymorphism at tens of thousands of different loci. A typical GWAS study will use more than one commercially available testing kit, thus increasing the number of the tested SNP loci to hundreds of thousands. The output of the testing of the cell sampling in this method is the unique genomic blueprint of the individual, including the specific DNA polymorphs his genome contains and/or the copy number of different genes (Figure 1).

Once these two biological databases are collected for each individual in the tested cohort, statistical tests are put to use. These tests are designed to identify, for each of the quantified phenotypes, whether any of the identified polymorphs in the genome correlates with the observed quantities of the specific phenotype. The tested cohort is partitioned into two according to the genotype possessed by each sample. If the subset *polymorph I* is shown to be of statistically significant lower transcript abundance than the subset *polymorph II*, then variation at this locus is taken to *correlate* with abundance levels of the specific RNA transcript.

Identification of a correlation of genomic polymorphism and RNA abundance levels at the statistical level is paralleled with *coidentification* at the conceptual level. The genomic polymorphism is considered to have an impact on the phenotype—thus becoming a functionally meaningful genotype. The phenotypic difference is shown to be affected by changes in the genome, thus becoming a genetically determined phenotype. In other words, it can be said that in the context of expression-based GWAS studies, correlation of the genotype and the phenotype is used to coidentify each other.

A second observation relates to the differentiation of the phenotypes into phenotypic classes. Once correlation of genomic polymorphism and RNA abundance levels is established, two distinct phenotypic classes are identified: *Genotype I* (formerly known as *polymorph I*) is identified with *low* RNA transcript levels; while *Genotype II* (*polymorph II*) is identified with *high* RNA transcript levels. The phenotype of RNA transcript abundance is, therefore, further split into two phenotypic classes of *high* and *low*, and a threshold value between *high* transcript abundance and *low* transcript abundance is identified.

In order to dispel possible confusion over the status of this observation, two cautionary statements are in place. The first relates to the "internal" validity of this finding: The split of the transcript level into *high* and *low* phenotype classes is a statistical observation on the tested population, *the study's cohort*. *Any single individual* in the tested population can be both part of the *low* phenotypic class and of *genotype II* (coidentified with the *high* phenotypic class), and vice versa. "Statistical observation" here only indicates that such occurrences are of significantly low probability in the sampled population.

The second caveat regards the external validity of such a finding on the general population: although the identification of threshold value between *high* and *low* transcript levels is of explanatory power for the sampled population, this is not necessarily so in *any other* population. Populations with different environmental conditions or genetic backgrounds can have different threshold values, or have this genotype masked by other influences.

These qualifications leave the main point intact: The statistically demonstrated correlation of DNA polymorphism and changes in abundance levels of an RNA transcript coidentifies the DNA polymorphs as genotypes, and the RNA transcript abundance as phenotypes. The coidentification of a G/P couple serves to establish the biological functionality of the specific DNA loci and the hereditary significance of the changes

in abundance level of the RNA transcript in the examined organism. The identification of hereditary traits in expression-based GWAS is not performed through identification of heredity patterns, but in direct correlation with variability in the genome itself. The completion of the Human Genome Project has made the identification of G/P couples using expression-based GWAS possible. This is a powerful tool for detecting genotype candidates that underlie differences between populations. Such couples may be related to differences in susceptibility to disease, effectiveness of treatment, or ethnically based differences.

Discussion

We have shown that, in current practice, geneticists use methods of statistical correlation between RNA abundance and DNA polymorphisms in order to identify genotypic and phenotypic differences. We now want to bring to the fore some interesting conceptual implications of this coidentification by statistical correlations, thus explicating some of the conceptual presuppositions of this technique. In particular, we shall call attention to two striking points:

(1) The expression-based GWAS analysis brings to the fore the coidentification of the genotype and the phenotype, suggesting a strong dependence between the two. We believe that this dependence is not a mere byproduct of the technique, but rather an indication of a conceptual dependence in the sense that the one is recognized as required for the individuation of the other.

(2) The coidentification of the genotype and the phenotype in this context emphasizes the strongly quantitative (and nonmechanistic) character of these terms in current molecular genetics—one which is based on correlations, and thus makes reference to a large population for its individuation and definition. In this sense, the current phenotype refers to its Johanssenian origins as a type of a population (Falk 2007), the difference being that in our molecular era the phenotypes are not macroscopic traits but microscopic features of specific tissues.

The identification of genome types in molecular genetics occurs in part through identification of phenome types rather than just by the DNA molecular sequence. Close examination of the techniques employed in the realm of expression-based GWAS suggests a substantial reliance on phenotyping in the way of identifying genotypes, and distinguishes the latter from mere polymorphism in the DNA. For molecular genetics, not every DNA variation underlies a different type. The identification of genome types is intimately tied to the identification of the entailed functionality (i.e., the identification of the phenome types). A DNA variation that does not produce a change in the phenome, at least in some circumstances, is not considered a functionally meaningful genetic variation.

While the identification of genome types requires the phenotype, the converse does not always hold. Any trait of an individual organism can in principle be identified as a phenotype, regardless of any underlying hereditary factor. Recall Lewontin's definition: "The phenotype of an organism is the class to which that organism belongs as determined by the description of the physical and behavioral characteristics of the organism, for example its size and shape, its metabolic activities and its pattern of movement" (Lewontin 2004). In light of the infinity of phenotypes this suggests, it is clear that there are phenotypes that are more biologically significant than others. Following Lewontin, we suppose here that some divisions of the realm of the individual organism's observed traits or *phenome* (Lewontin 2004) are more significant than others, and therefore, there are phenotypes that are more significant than others. This significance of the phenotype is a result not only of its own nature but also of the way we discern and measure it. It is the technology of choice that sets the nature of what is actually measured, which in turn plays a cardinal role in determining the nature of those phenotypes that will be deemed biologically significant. In the case at hand, the technology used measures the abundance of each and every RNA transcript in an individual tissue sample, treating it as a quantitative phenotype. Rather than identifying stand-alone phenotypes, expression-based GWAS identify those from the pool of RNA transcript abundance phenotypes that are differentially regulated by the DNA sequence, forming a subdivision of genetically regulated quantitative phenotypes.[4] Independently of the mechanism of operation underlying the genetic basis of the phenotype, the statistical method employed by GWAS allows for statistically significant correlation and coidentification of the genotype with a subset of phenotypic classes.

Recent association studies on the transcriptome highlight the significant role of RNA abundance in affecting the traits of the organism. A study on adipose tissue of more than 5,000 members of Icelandic families suggests the possibility that phenotypic traits in the organism are brought about by the amount of the RNA transcript as much as by the RNA base sequence (Emilsson et al. 2008). A key finding of this study is that more than 50% of all RNA transcripts are strongly correlated with clinical traits related to obesity. Through linkage analysis, this same study supports the "extensive genetic component underlying gene expression traits in . . . adipose tissue" (Emilsson et al. 2008: 423). Other association studies support similar findings.[5] In 2007, the U.S. Food and Drug Administration approved the first clinical diagnosis based on measurement of RNA transcript abundance.[6]

These studies underlie a shift in perception that is taking place in contemporary genetics: Whereas abundance of the RNA was once considered as a factor controlled by cellular mechanisms, the emerging perception from recent GWAS studies suggest that genome sequence and structure play an extensive, nontrivial role in the regulation of RNA transcript abundance. In turn, the control of RNA abundance levels affects the phenotypic traits of the organism at both the micro- and macro levels. To use functional language, the genome sequence has the role of providing both a template for the transcription of RNA and regulation of abundance of this transcript.

GWAS clearly still provide the ground for the generation of hypotheses over possible mechanisms.[7] Having said that, the mechanism of operation is not an essential part of the conceptualization or the modeling in expression-based GWA studies, but rather relegated to the periphery of the description, interpretation, and mitigation of the findings. A molecular-mechanism explanation does not make redundant the explanatory force of the functional ascription in the G/P couple, but rather elaborates the way this function is executed *at that instant* in the same way as elaborating on the mechanism of operation of the heart does not make redundant the explanatory force of the functional ascription of the heart as a blood pump in the circulatory system.

Expression-based GWAS use the abilities of microarray technology and SNP libraries to perform wide population studies. As the identification of the G/P couple is through statistical significance association, many G/P couples, falling short of significance threshold, are not identified. Indeed a feature of expression-based GWAS is that it identifies G/P couples in a certain population over certain environmental conditions. Thus, expression-based GWAS performed on different populations show that certain G/P couples are concealed in each of the populations (cf. Spielman et al. 2007). An intriguing question for further research is whether the identification of G/P couples can be extended beyond the investigated population, or whether such coupling is dependent on environmental conditions. This raises the possibility that the genotype itself, as conceived by GWAS, is not independent from the environmental conditions met by the study population.

In closing, let us remark that we find some irony in the fact that the method of molecularization employed in current GWAS refers us back to the conceptual foundations of classical genetics in at least two respects: Mendel's method of inferring elementary factors from visible traits, and Johannsen's distinction between the phenotype and the genotype through the use of population studies have undergone molecularization. More than ever before, molecular genetics seems to be revealing newer and finer manifestations of its classical roots by means of rapidly improving technology. Thus understanding the future of molecular biology could be enhanced by reflection on its past.

Acknowledgments

Research on this article was supported by a generous grant from the ISF (345-06). Early versions of this article were presented at the annual meetings of the Israeli Association for the History and Philosophy of Science (April 2009), and we would like to thank all participants in the discussion and in particular Raphael Falk and Oren Harman for comments and suggestions. We would also like to thank Oren Harman for reading and commenting on a late version of the article. Critical comments by a referee for this journal have significantly contributed to the final version.

Notes

1. Sterelny and Griffiths (1999) conceptualized Morgan's approach for a gene as a *phenotype difference maker*.

2. Griffiths and Stotz (2007) use the term "nominal gene" in reference to a similar idea of a structural stereotype or a prototype of a gene.

3. The mutual dependence thesis in this article is not intended to contribute to the well-known discussion on the levels of selection, nor do we ask whether the general relation between the level of the genotype and the level of the phenotype is to be conceptualized in terms of reduction to the lower level (Sterelny and Kitcher 1988), upper level (Griesemer 2003; Callebaut 2005), or pluralistically rather than reductionistically (Lloyd 2005).

4. In a previous article (Nachtomy et al. 2007), we have suggested the name "endophenotypes," employed in psychiatry for measurable components unseen by the unaided eye between disease and genotype, to denote this subclass of phenotypes.

5. For pioneering work on this subject, see Golub et al. (1999) and Bittner et al. (2000).

6. The device known as MammaPrint was cleared by the U.S. Food and Drug Administration (2007). The results of the clinical study conducted by this device were published in van't Veer et al. (2002).

7. In Rockman and Kruglyak (2006), the authors discuss possible mechanisms of operation linking the RNA coding regions with the genotypic site.

References

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406: 536–540.

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752–755.

Callebaut W (2005) Again, what the philosophy of biology is not. Acta Biotheoretica 53: 93–122.

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO (2007) A genome-wide association study of global gene expression. Nature Genetics 39: 1202–1207.

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson

T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K (2008) Genetics of gene expression and its effect on disease. Nature 452: 423–428.

Falk, R (2007) Wilhelm Johannsen: A rebel or a diehard? In: Rebels, Mavericks, and Heretics in Biology (Harman O, Dietrich MR, eds), 65–83. New Haven, CT: Yale University Press.

Fogle T (2001) The dissolution of protein coding genes in molecular biology. In: The Concept of the Gene in Development and Evolution: Historical and Epistemological Perspectives (Beurton RF, Falk R, Rheinberger HJ, eds), 3–25. Cambridge, UK: Cambridge University Press.

Goldstein, D (2009). Common genetic variation and human traits. New England Journal of Medicine 360: 1696–1698.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286: 531–537.

Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature Genetics 39: 1208–1216.

Griesemer JR (2003) Genetics from an evolutionary process perspective. In: Genes in Development: Re-reading the Molecular Paradigm (Neumann-Held EM, Rehmann-Sutter C, eds), 343–375. Durham, NC: Duke University Press.

Griffiths PE, Stotz K (2007) Gene. In: The Cambridge Companion to the Philosophy of Biology (Ruse M, and Hull D, eds), 85–102. Cambridge, UK: Cambridge University Press.

Hedenfalk I, Ringner M, Ben-Dor A, Yakhini Z, Chen Y, Chebil G, Ach R, Loman N, Olsson H, Meltzer P, Borg A, Trent J (2003) Molecular classification of familial non-BRCA1/BRCA2 breast cancer. Proceedings of the National Academy of Sciences USA 100: 2532–2537.

Hirschhorn J (2009) Genomewide association studies: Illuminating biologic pathways. New England Journal of Medicine 360: 1699–1701.

Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G, Schmidt S, Müller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, Aitman TJ (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nature Genetics 37: 243–253.

Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology 3: 318–356.

Keller EF (2002) Making Sense of Life. Cambridge, MA: Harvard University Press.

Lewontin RC (1992) Genotype and phenotype. In: Keywords in Evolutionary Biology (Keller EF, Lloyd E, eds), 137–144. Cambridge, MA: Harvard University Press.

Lewontin RC (2004) The genotype/phenotype distinction. In: Stanford Encyclopedia of Philosophy. Avaliable at http://plato.stanford.edu/entries/genotype-phenotype

Lloyd EA, Dunn M, Cianciollo J, Mannouris C (2005) Pluralism without genic causes? Philosophy of Science 72: 334–341.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430: 743–747.

Nachtomy O, Shavit A, Yakhini Z (2007) Gene expression and the concept of the phenotype. Studies in History and Philosophy of Biology and Biomedical Sciences 38: 238–254.

Office of Population Genomics (2009) Catalog of Published Genome-Wide Association Studies. National Genome Research Institute, National Institute of Health. Available at http://www.genome.gov/gwastudies

Osada N, Kohn MH, Wu CI (2006) Genomic inferences of the cis-regulatory nucleotide polymorphisms underlying gene expression differences between *Drosophila melanogaster* mating races. Molecular Biology and Evolution 23: 1585–1591.

Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nature Review Genetics 7: 862–872.

Roeder K, Luca D (2009) Searching for disease susceptibility variants in structured populations. Genomics 93: 1–4.

Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, Mayford M, Lockhart DJ, Barlow C (2000) Regional and strain-specific gene expression mapping in the adult mouse brain. Proceedings of the National Academy of Sciences USA 97: 11038–11043.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G, Linsley PS, Mao M, Stoughton RB, Friend SH (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297–302.

Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG (2007) Common genetic variants account for differences in gene expression among ethnic groups. Nature Genetics 39: 226–231.

Sterelny K, Griffiths PE (1999) Sex and Death: An Introduction to Philosophy of Biology. Chicago, IL: University of Chicago Press.

Sterelny K, Kitcher P (1988) The return of the gene. Journal of Philosophy 85: 339–360.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.

U.S. Food and Drug Administration (2007) Letter 510(k) Summary: K070675. Available at http://www.fda.gov/cdrh/pdf7/K070675.pdf

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536.

Waters KC (2000) Molecules made biological. Revue Internationale de Philosophie 214: 9–34.

Weber M (2005) Philosophy of Experimental Biology. Cambridge, UK: Cambridge University Press.