Introduction

A human cell is defined by its components, such as the genome, epigenome, proteome, metabolome or transcriptome, and their interactions. This results in a complex regulatory network that we just begin to understand and that poses a major challenge in finding the cellular cause of a given human disease. Even though a systems biological approach integrating all aspects that define a cell type would be best suited to understand human development and disease, researchers only slowly start to leave the isolation of their own specialized -Omic domain.

The field of genomics is likely the most advanced in its global search for disease-associated alterations of the genome. Already for decades, inheritance studies based on genetic linkage in families have been used to map genomic loci that have an effect on disease or other phenotypic traits. Linkage analysis relies on the co-segregation of marker alleles, which are, for example, common single nucleotide polymorphisms (SNPs) with the unknown disease gene within pedigrees. While this approach has had great success for diseases and traits that are controlled by a single locus (Mendelian traits) (Botstein and Risch 2003), it has proven cumbersome for the analysis of common and complex diseases such as cancer (Aldrich et al. 2001). Already in 1996, Risch and Merikangas proposed the performance of an association scan that involves millions of common variants of the genome with the unknown disease gene within pedigrees. This approach has had great success for diseases and traits that are controlled by a single locus (Mendelian traits) (Botstein and Risch 2003), and it has proven cumbersome for the analysis of common and complex diseases such as cancer (Aldrich et al. 2001). Already in 1996, Risch and Merikangas proposed the performance of an association scan that involves millions of common variants of the genome and a group of unrelated individuals that differ in a certain phenotype. In particular for complex traits this approach should yield much better results than a linkage analysis including only a few hundred markers (Risch and Merikangas 1996). Based on this principle, the first genome-wide association study (GWAS) published in 2005 (Klein et al. 2005) marks the beginning of a whole new era of research counting 1,600 published...
GWA reports and 10,088 disease-associated SNPs by May 2013 (Hindorff LA 2013).

Even though bearing great promise, the success of GWAS for clinical benefits such as the discovery of new biomarkers that can be used for clinical decision support or disease prevention remains limited. There are two main reasons for this: First, the problem of missing heritability and second, the limited identification and functional characterization of causal variants. Heritability is defined as the proportion of the phenotypic variance in a population that is due to genotypic differences among individuals (Gibson and Shepherd 2012). For example, human height has an estimated heritability of 80 %, meaning 80 % of height differences between individuals can be explained by genetic differences and 20 % are due to other influences such as the environment. Even though 40 genomic loci have been identified to be associated with human height, they only explain 5 % of the phenotypic variance (Visscher 2008). Multiple reasons have been suggested to explain the missing heritability, one of them being the fact that GWA studies typically identify common variants (present in 5 % or more of the population) with small effects and miss out on rare variants (allele frequency <1 %) with potentially much larger effects. This topic is extensively reviewed in Manolio et al. (2006) and Gibson (2011).

In this review, we will focus on the second aspect: The identification and, in particular, the functional characterization of causal variants. Principles for the post-GWAS functional characterization of risk loci are also reviewed elsewhere (Freedman et al. 2011); however, possibilities that mass spectrometry-based proteomics can offer are not discussed. We will summarize both (epi-)genomics and proteomics technology in light of post-GWAS, and thus hope to provide a basis for an highly integrated, systems biological approach. As we cover multiple broad topics, we apologize that due to space restrictions we were not able to cite all relevant publications and would like to refer to other reviews cited in the text. Throughout the review, we mainly discuss SNPs or small genomic variants, but we recognize that other types of common genetic variation, such as larger insertions or deletions, may also influence risk.

Identification of all common and rare associated variants

If a certain combination of genomic loci in a population occurs more or less often than would be expected from a random formation, they are defined to be in linkage disequilibrium (LD) with each other (reviewed in Slatkin 2008). It is a second-order phenomenon derived from linkage, which is the presence of two or more loci on a chromosome with limited recombination between them. SNPs represented on a GWAS SNP array were chosen in a way that they capture the LD structure of the genome and thus allow the identification of associations between a common trait and a certain genomic region that is represented by one marker (tagSNP). Hence the associated SNP is not always the causal variant and any other SNP or a combination of SNPs that are in strong LD with the tagSNP can form the basis of functional consequences. For this reason, one of the major tasks of the HapMap project is to identify all common and rare variants to generate a comprehensive catalog of human genome variations (Altshuler et al. 2010).

In 2012, the 1,000 genomes project consortium published genomes of 1,092 individuals from diverse ethnic populations using a combination of low-coverage whole-genome and exome sequencing (Genomes Project Consortium 2012). This study captures up to 98 % of accessible SNPs that have a frequency of 1 % or higher in UK-sampled genomes and with 38 million SNPs and approximately 1.5 million other variants provides an extensive resource of common and rare variants. Trait-associated SNPs that are in LD with a certain tagSNP can thus be imputed (Howie et al. 2012). While this approach shows great success for common variants (>5 % frequency), rare variants are more recent and thus geographically restricted. For this reason, many more individuals from different populations around the globe need to be sequenced to provide good coverage. In addition, large efforts are currently taken to fine map regions that are associated with a certain disease phenotype by extensive targeted re-sequencing (Nejentsev et al. 2009; Rivas et al. 2011).

Identification and characterization of the functional variants

Having successfully identified thousands of novel common and rare variants that are in LD with previously characterized GWAS tagSNPs, the next big challenge is to find the causal variants amongst those. Most methods that have been developed so far focus on SNPs that are located in the coding or transcribed region of a gene because these might influence the primary structure and thus the function of a protein (Ng and Henikoff 2003; Saccone et al. 2011; Cvijic et al. 2013). However, most of the associated common variants identified so far do not map within or in LD to a protein coding region (Easton et al. 2012) and thus might be rather linked to gene expression regulatory mechanisms. Their characterization remains difficult and requires the integration of data from related fields such as epigenetics or proteomics.
Integration of GWAS with epigenetics information on regulatory elements

A SNP located in a non-coding region may, for example, disrupts or creates a transcription factor (TF)-binding site in an active regulatory element (Reddy et al. 2012). As a consequence, the regulatory activity and thus the expression of a gene that is controlled by this element can be altered (Kasowski et al. 2010). A GWAS SNP that overlaps with an active regulatory region or an experimentally detected TF-binding site in a relevant cell type therefore has a higher probability of being functionally relevant (Jia et al. 2009; Harismendy et al. 2011; Paul et al. 2011).

In fact, a recent study involving genome-wide DNase I mapping in 349 cell and tissue samples showed that 76.6 % of all non-coding GWAS SNPs either lie within a DNase I hypersensitive site (DHS) or are in complete LD with SNPs in a nearby DHS (Maurano et al. 2012). Besides studying histone modifications and the binding pattern of TFs by chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Johnson et al. 2007; Robertson et al. 2007), DNase I hypersensitive site identification by sequencing (DNase-seq) or digital genomic footprinting (DGF) (Crawford et al. 2006; Boyle et al. 2008; Hesselberth et al. 2009) are major techniques to map regulatory elements (Visel et al. 2009). We recently used the combination of these technologies to define binding sites and thus the regulatory impact of the oncprotein PML-RARα and AML1-ETO in acute myeloid leukemia (Saeed et al. 2012).

Large epigenetics consortia such as ENCODE (http://genome.ucsc.edu/ENCODE/), Roadmap (http://www.roadmapepigenomics.org/), iHEC (http://www.ihec-epigenomes.org/) or BLUEPRINT (http://www.blueprint-epigenome.eu) utilize next-generation sequencing technology to characterize, amongst others, histone modifications, TF binding and chromatin accessibility in various cell types, human tissue and blood, respectively. This enormous resource can be used to characterize the regulatory landscape of susceptibility regions in relevant cell types and narrow down causal variants to those mapping to an active regulatory element. RegulomeDB is a database that combines data from ENCODE and other sources such as ChIP-seq data from the sequence read archive (SRA) (Leinonen et al. 2011) and data on expression quantitative trait loci (eQTLs) with computational predictions to estimate the regulatory potential of a certain genomic locus (Boyle et al. 2012). A parallel publication of the same group shows that SNPs that are annotated with a high score are in most occasion SNPs that are in LD with a reported association rather than the tagSNP itself (Schaub et al. 2012b).

Identification of physiological relevant target genes

As mentioned above, one of the major mechanisms underlying susceptibility to complex trait or disease is probably the variation in gene expression caused by polymorphisms in regulatory elements. Consequently, transcript abundance can be analyzed with genetic approaches in the same way as any other quantitative trait phenotype, such as height or the body mass index and are commonly known as expression quantitative trait loci (eQTLs) (reviewed by Cookson et al. 2009). A SNP in a non-coding genomic region could thus be linked to a certain phenotypic trait in a GWAS study and the same SNP or a SNP in strong LD could be linked to expression changes in an independent eQTL study, thus providing an important connection between a phenotypic trait and a physiological relevant target gene.

Mapping eQTL target gene associations in tumors is more challenging than for other human traits or disease. Tumors acquire frequent genetic and epigenetic alterations, which can substantially affect gene expression (Raval et al. 2007; Smith et al. 2006) and consequently obscure the association between germline genetic polymorphisms and gene expression (Curtis et al. 2012). For these reasons, recent cancer studies also investigate the association between SNPs and an altered epigenetic landscape, such as promoter methylation, histone modifications or the expression of large intergenic non-coding RNAs (lincRNAs) that associate with chromatin-modifying complexes (Gibbs et al. 2010; Bell et al. 2011; Grossman et al. 2013; Ernst et al. 2011). A good example is the recent work from Li et al. (2013), which provides a more comprehensive picture of gene expression determinants in breast cancer and the underlying biology of breast cancer risk loci by the integrated analysis of eQTLs, somatic copy number alteration and CpG methylation in 219 tumor samples and the healthy counterparts.

Even though quantitative trait loci analysis can indicate an impact of a genomic region on the expression or the epigenetic regulation of certain genes, it is unclear if this influence is direct or indirect. Mammalian genomes are organized into higher-order conformational structures that allow physical interactions of regulatory elements that can be located in far distance on one chromosome or even on different chromosomes (reviewed in Cremer and Cremer 2001). Chromosome Conformation Capture (3C) and similar techniques have been developed to identify these interactions and demonstrated their impact on the regulation of transcriptional and epigenetic states (reviewed in de Wit and de Laat 2012). Hi-C, for example, allowed the investigation of the three-dimensional organization of the human and mouse genomes in embryonic stem cells and terminally differentiated cell types at unprecedented resolution (Dixon et al. 2012). Chromatin Interaction Analysis.
by Paired-End-Tag sequencing (ChIP-PET) is a complementary methodology that is used for the genome-wide mapping of chromatin interactions bound by specific proteins (Fullwood et al. 2009). Applying this technology to proteins generally bound by promoters or enhancers (e.g., PolII) allows the high-resolution mapping of enhancer-promoter and promoter–promoter interactions (Li et al. 2012). Data on higher-order conformational structure of a relevant cell type can thus help to unambiguously identify direct physiological target genes of functional GWAS SNPs.

Prediction and validation of SNP-dependent differential transcription factor binding

Integrated studies often use TF-binding motifs present in DHS or DNase I footprints and overlapping ChIP-seq peaks to predict differential TF binding caused by an SNP (Schaub et al. 2012a; Maurano et al. 2012). While this is a great approach to restrict the number of SNPs associated with a certain phenotype to those that might have a causal role, like any generic approach it can also reveal a number of false positives and negatives. False hits might be the result of using databases with data generated from multiple, often for a disease or trait phenotype not relevant cell types. Many distal regulatory elements are cell type specific (Heintzman et al. 2009; Dimas et al. 2009) and thus a transcription factor that binds to a region with an SNP might not even be expressed in another cell type relevant for a certain trait or disease. Furthermore, the haplotype for the particular SNP of the cell lines used in the database is often not taken into account for these analyses. Despite huge efforts that have been taken to characterize TF-binding motifs (Badis et al. 2009; Jolma et al. 2013; Noyes et al. 2008), only for approximately half of the more than 1,000 human TFs a corresponding DNA binding motif is known, thus introducing a bias towards those. Last but not least, most motif-based approaches consider TF-binding motifs in an isolated context. However, several TFs that are part of one TF family might compete for the same motif and TFs binding in close proximity might influence each other’s affinities. For these reasons, prediction-based methods cannot yet replace the biochemical characterization of differential TF binding and activity.

Electromobility shift assays (EMSAs) are the classical method to study the interaction between a certain protein or protein domain with a particular sequence of DNA (Fried and Crothers 1981; Garner and Revzin 1981). This in vitro method, however, requires pure or highly enriched protein and, in an ideal case, an antibody that is specific for the studied TF. Chromatin immunoprecipitation (ChIP) followed by quantitative PCR of the region containing the SNP is probably the method of choice to validate the predicted differential TF binding in vivo (Stunnenberg and Vermeulen 2011). However, two major requirements need to be met. First, sufficient amounts of two disease relevant cell or tissue types that are homozygous for either variant of the SNP or alternatively a heterozygous cell type are needed. Second, a ChIP-grade antibody that recognizes the TF in question needs to be available. Furthermore, this method is a validation method that requires a priori knowledge of the binding TF. The predictions mentioned above might reveal multiple candidates each requiring a separate validation experiment. Last but not least, ChIP experiments cannot distinguish between directly and indirectly bound proteins, and without performing additional, hypothesis-driven experiments it will not reveal information about recruited co-factors and protein complexes, which would significantly contribute to our understanding of the underlying regulatory mechanism.

In the following, we will introduce recent developments in the field of mass spectrometry-based proteomics that, amongst others, will be able to overcome at least some of the obstacles mentioned above.

Mass spectrometric characterization of functional SNPs or variants

Due to numerous technical and computational developments during the last decade, the detection and quantification of proteins in complex mixtures by mass spectrometry have evolved to a standardized methodology (reviewed in Ahrens et al. 2010). Besides the analysis of complete proteomes (de Godoy et al. 2008) and the quantification of post-translational modifications (reviewed in Choudhary and Mann 2010), the technique has also been widely used to study protein interactions and complexes in an unbiased manner (reviewed in Vermeulen et al. 2008), thus offering an alternative to affinity purification followed by western blotting with specific antibodies. Here, we will review recent developments in the proteomics field that can be perfectly integrated in post-GWAS studies and thus contribute significantly to the identification and functional characterization of trait-associated SNPs or variants (Fig. 1).

Integration of comprehensive protein expression data

In contrast to transcriptomics, proteomics for long had the disadvantage of not being comprehensive and requiring substantial amounts of material. The limited sequencing speed of mass spectrometers as well as the immense dynamic range of the human proteome made it difficult to identify all proteins in a reasonable time frame and with reasonable effort. Recent developments of novel methods, software and instrumentation now allow the identification of comprehensive proteomes as demonstrated for yeast a couple of years ago (de Godoy et al. 2008). In minimal amounts
of human cells or tissue more than 10,000 proteins can be detected, presumably covering most of the expressed proteins (Wisniewski et al. 2013; Munoz et al. 2011). In contrast to next-generation sequencing, in which the number of reads for a certain genomic region is directly proportional to the amount of DNA or RNA present in the sample, mass spectrometry is not inherently quantitative. As a result, the detected peptide and consequently protein intensities do not represent the absolute abundance of proteins in a cell. Internal standards and computational normalization can solve this issue (Schwanhäusser et al. 2011; Picotti et al. 2009; Zeiler et al. 2012). In summary, it is now possible to obtain comprehensive proteomes with copy number information for each protein from a minimal amount of material.

As described above, approaches such as the one used in RegulomeDB integrate published epigenomic datasets from multiple cell lines to indicate which SNPs are likely to have a functional influence. In an ideal case scenario, this analysis would be done on data generated in a disease or trait relevant tissue. While large efforts are being undertaken to map epigenetic marks and DNA hypersensitivity in most human tissues (Adams et al. 2012; Chadwick 2012), ChIP-seq profiles of all TFs in all tissues are unlikely to be available in the near future. Comprehensive proteomes provide information about the presence and abundance of TFs in a certain cell type. Therefore, proteomic profiles of all tissues can serve as a filter for TF-binding predictions based on DNA motifs or ChIP-seq profiles in common cell lines. We thus strongly propose to include comprehensive and quantitative proteome mapping into large-scale epigenome mapping efforts.

Protein quantitative trait locus analysis

It is known for more than a decade that global mRNA and protein levels do not correlate well (Gygi et al. 1999b). Reasons for this are various layers of post-translational regulation that buffer changes in transcript abundance or lead to alterations in protein abundance despite a constant transcript level. This raises the question whether polymorphisms in eQTLs have a comparable effect on transcript and protein levels.

Already in 2007, the first protein quantitative trait locus (pQTL) study was performed, analyzing the proteomes of...
two laboratory yeast strains and 98 segregants (Foss et al. 2007). All of these strains had also been genotyped and studied with regard to the genetic basis of variation in transcript levels (Brem et al. 2005). From this study, it became clear that loci that influence protein levels differ from those that influence transcript abundance. This emphasizes the importance of the direct analysis of the proteome. However, the proteomics technology at that time yielded limited proteome coverage. Furthermore, a very small set of proteins was quantified across all samples resulting in a strong bias towards high abundant proteins. Targeted mass spectrometry now allows the reliable quantification of a selected set of approximately 50 proteins across a broad abundance range and a large number of samples (Picotti et al. 2009). On this basis, the study of 2007 was repeated, resulting in a much more complete dataset while at the same time requiring less measurement time (Picotti et al. 2013). Novel pQTLs as well as epistatic interactions were detected. Furthermore, the authors identified two cases of co-inheritance of several independent genetic variations that influence the abundance of related proteins in a biologically coherent manner. Recently, the first human pQTL study comparing the proteome of lymphoblastoid cell lines from 95 individuals that were genotyped in the HapMap Project (www.hapmap.org) was published (Wu et al. 2013). Similar to the yeast study mentioned above, the authors stressed the limited overlap between eQTLs and pQTLs, indicating that distinct and diverse genetic mechanisms control gene expression at many different levels.

A big bottleneck in these studies, however, is the limited throughput. As the samples were measured consecutively, a still substantial amount of measurement time was required. Recent developments on the basis of isotope labels, introduced either metabolically in cell culture and/or chemically on peptide level (for detailed review, see Nikolov et al. 2012), allowed the measurement of up to 54 samples in a single experiment (Everley et al. 2013; Hebert et al. 2013). While metabolic labeling such as stable isotope labeling by amino acids in cell culture (SILAC) (Ong et al. 2002) is a very elegant, highly accurate method, it is only applicable to cells dividing in culture and thus cannot be used for the multiplexed analysis of human tissue samples. The development of multiplexing approaches such as the ones mentioned above that solely rely on chemical labeling (Boersema et al. 2009; Ross et al. 2004; Thompson et al. 2003; Gygi et al. 1999a) would therefore immensely benefit the feasibility of human pQTL studies.

Integrating large-scale data on protein–protein interactions to interpret GWAS

In recent years multiple efforts, mainly employing yeast two-hybrid screens or affinity purification followed by mass spectrometry (AP-MS) have been undertaken to construct high confidence interaction networks of human proteins (Venkatesan et al. 2009; Glatter et al. 2009; Kristsens et al. 2012; Hubner et al. 2010; Mellacheruvu et al. 2013; Sowa et al. 2009). Most AP-MS approaches are based on cell lines that express tagged versions of the proteins of interest and employ quantitative methods, such as the ones described in the following section for DNA-pulldowns, to ensure high specificity (Fig. 2). Significant efforts and resources are required to reach large-scale interaction networks including a large number of proteins. For this reason, a comprehensive interaction network including all proteins that are expressed in a certain cell of interest has still not been published.

Based on published protein–protein interaction data, there have been successful attempts to integrate these networks with GWAS or traditional linkage studies to assist the prioritization of candidates genes as well as to provide a possible functional background (Califano et al. 2012). For example, Lage and co-workers created a phenome–interactome network of protein complexes implicated in genetic disorders. Based on this, they provide numerous novel disease-causing candidate genes implicated in various diseases such as inflammatory bowel disease or Alzheimer (Lage et al. 2007). Another group developed a high confidence algorithm to in silico predict protein–protein interactions that are not yet covered by the experimental procedures mentioned above (Elefsinioti et al. 2011). Subsequently, this ‘comprehensive’ protein–protein interaction network was applied to study molecular mechanisms of neurodegenerative diseases by integrating it with relevant GWAS. This analysis provided evidence of the involvement of TOMM40 in Alzheimer’s diseases.

DNA-pulldowns to identify and study dynamic DNA–protein interactions

Quantitative proteomics allows the study of dynamic interactions of proteins or entire protein complexes with a certain DNA sequence (Fig. 2) (Ranish et al. 2003; Mittler et al. 2009). A biotinylated, double-stranded oligonucleotide of approximately 30 base pair length containing a TF-binding motif of interest is immobilized on streptavidin beads. In parallel, the same oligonucleotide containing a point mutation in the motif is immobilized on a second set of beads serving as a control. Both DNA fragments are incubated with differentially SILAC labeled nuclear extract from a cell line of interest. In several, low stringent wash steps, unbound proteins are removed and the experiments are merged into a single tube. The DNA and the bound proteins are released from the beads by cleaving with a restriction enzyme specific for a recognition sequence included into the bait oligo sequence. Alternatively, oligonucleotides
Fig. 2 Illustration of different DNA-pulldown workflows. General workflows for the identification of SNP-dependent, dynamic DNA–protein interactions using DNA-pull-downs based on metabolic isotope labeling (a), chemical labeling (b) or label-free protein quantification (c). a A cell line of interest is cultured in medium containing light (C12N14) or heavy (C13N15) arginine and lysine. After full incorporation of amino acids into the proteome, nuclear extracts are prepared. Biotinylated oligonucleotides containing either variant of the SNP are immobilized on streptavidin beads and incubated with the light and heavy nuclear extract. Unbound proteins are removed by several wash steps. Subsequently, proteins are eluted and differentially labeled eluates are mixed prior to tryptic digestion. Peptides are separated and identified using reversed phase liquid chromatography coupled online to a mass spectrometer (LC–MS/MS). SILAC ratios from two replicate experiments are plotted against each other. Dynamic SNP interacting proteins (large or small ratio) can thus be distinguished from unspecific background binders or proteins that bind to other parts of the oligonucleotides (log2(ratio) = 0) (Mittler et al. 2009). b In contrast to the workflow described in a, a normal, unlabeled nuclear extract from cells or tissue can be used. After the pulldown, proteins are eluted and digested separately. Subsequently, peptides from both pulldowns are differentially labeled by chemically introducing isotopes at the N-termini and the arginine and lysine side chains (Ranish et al. 2003). c In the label-free approach, all steps, including the LC–MS/MS acquisition, are carried out separately. Peptide intensities between all runs are compared using advanced label-free quantification algorithms. Dynamic SNP interacting proteins can be identified by their differential peptide intensities and their p value in a t test based on triplicates (Hubner et al. 2010).
can be tagged with desthiobiotin and eluted with biotin (Butter et al. 2012). Eluates are digested and analyzed by mass spectrometry. The SILAC ratios will allow the discrimination of proteins that have a higher affinity to one of the oligonucleotides and thus to one of the SNP alleles. As mentioned above, SILAC labeling is only applicable when using extract from cells that can be grown in culture. Alternatively, chemical labeling of peptides or label-free protein quantification can be used (Hubner and Mann 2011). Recently, this method allowed the identification dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives, an important resource for the field of epigenomics (Spruijt et al. 2013; Bartels et al. 2011). Furthermore, a variation of this approach was developed to identify proteins that specifically interact with an RNA stem-loop of interest (Scheibe et al. 2012). In addition, complementary techniques allow the identification of proteins that are associated with a single, in vivo crosslinked and purified genomic locus (Dejardin and Kingston 2009; Byrum et al. 2012). However, due to the enormous amounts of cell material that is required for these methods it remains extremely difficult and thus is not discussed here any further.

Even though DNA-pulldowns could significantly contribute to our understanding of the molecular consequences underlying human genetic variability, to our knowledge this has so far only been shown by two studies. First, DNA-pulldowns were used to identify a repressor protein, muscle growth regulator (MGR), that specifically binds to an SNP in the intron of IGF2 and leads to enhanced muscle growth in European pigs (Butter et al. 2010). Recently, a second study that applies DNA-pulldowns to study allele-specific TF binding to SNPs that are highly associated with type 1 diabetes has been published (Butter et al. 2012). The authors could reduce the initial set of 12 associated SNPs at the IL2RA locus to four that showed differential binding of TFs and thus might have a functional impact on the disease. The limited interaction between the GWAS and the proteomics community might be the major reason for the minimal employment of this approach in post-GWAS studies.

An intrinsic problem is the throughput of DNA-pulldowns. As mentioned above, each GWAS study might reveal hundreds of rare or common SNPs that are in LD with an SNP that is linked to a certain disease. The method described above, however, is only capable of profiling 5 SNPs per day and mass spectrometer. Furthermore, at least 800 µg of nuclear extract is needed for each experiment. These experiments should ideally be performed using suitable (i.e. diseased) primary material to make sure that relevant proteins are expressed at correct levels and in the correct state (splice isoforms, post-translational modifications). However, for most primary cells and tissues it will not be possible to provide the large amount of material that would be necessary for a larger screen. We are convinced, however, that the recent development of integrated computational approaches, that limit the possible functional candidates in a set of associated SNPs, as well as efforts in downscaling and potentially multiplexing DNA-pulldowns will close the gap in near future.

While DNA-pulldowns so far have only been described to map differential binding of TFs to a specific sequence containing one or multiple nucleic acid variants, it can in theory also be used to identify the sum of TFs and cofactors that bind to a specific locus of interest. This could be achieved by quantitatively comparing the protein abundance across pulldown experiments from different genomic regions. This concept was recently used in a targeted approach to identify proteins binding to the FLO11 promoter region (Mirzaei et al. 2013). An alternative could be a discovery-based mass spectrometric acquisition method combined with a label-free quantification algorithm. We and others have already shown this concept for protein–protein interactions (Sowa et al. 2009; Hubner et al. 2010).

Concluding remarks

GWAS studies provide important information about associations between phenotypic traits and genomic loci. Now, in the post-GWAS era, a major task is to decipher the biological processes and functional mechanisms underlying these associations. This requires the rigorous integration of large-scale, multi-dimensional data and expertise from various fields. Numerous, fruitful collaborations have already been established between researchers in genetics, genomics and epigenomics. These integrated analyses are “straightforward” as they rely mainly on a similar technology platform and data output from next-generation sequencing. Other fields studying the proteome or the metabolome rely on mass spectrometric measurements and thus completely different experimental set-ups and analysis pipelines. This might be the major reason why the possibilities that proteomics research offers are so far hardly recognized and integrated in post-GWAS studies.

In this review, we summarized the current ‘post-GWAS workflow’ involving the identification of variants that are in linkage disequilibrium with a GWAS variant, the identification of the functional variant among those, the identification of physiological target genes as well as the characterization of the biological mechanism underlying the functional variant. Previous studies that showed the limited correlation of mRNA and protein levels in a cell (Gygi et al. 1999b) as well as the incomplete overlap of expression and protein quantitative trait loci (Foss et al. 2007; Wu et al. 2013) stress the importance of expanding the portfolio of resources that are currently used for GWAS follow-up. Therefore, we introduced recent developments in the field
of proteomics and suggested how those can be efficiently integrated in the workflow outlined in Fig. 1. For example, DNA-pulldowns followed by mass spectrometry allow the unbiased characterization of SNP-dependent protein-DNA interaction dynamics such as the altered recruitment of transcription factor complexes (Butter et al. 2012).

Clearly, the road towards fully integrative, quantitative biology to study the functional mechanisms of GWAS SNPs does not end with the integration of proteomics. The system-wide profiling of protein post-translational modifications, for example phosphorylation, glycosylation or acetylation, as well as the influence of thousands of metabolites on the phenotypic appearance of a cell provide additional, very powerful datasets that could be integrated in current post-GWAS workflows. In addition, the results obtained by the approaches described in this review will need to be followed up by extensive, more detailed functional studies involving cell and tissue models to further unravel the pathogenesis underlying a certain disease trait. Taken together, GWAS could be the basis for so far unseen collaborative efforts that provide new directions for the prevention and treatment of common diseases.

Acknowledgments We thank M. Vermeulen for critical reading of the manuscript. This review was supported by the EU FP7 framework program SYSCOL (systems biology of colorectal cancer). Nina C. Hubner was supported by a Veni grant from NWO, The Netherlands.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

A Catalog of Published Genome-Wide Association Studies (2013). http://www.genome.gov/gwastudies. Accessed 1 May 2013.

Adams D, Altucci L, Antonarakis S, Ballesteros J, Beck S, Bird A, Ahrens CH, Brunner E, Qeli E, Basler K, Aebersold R (2010) Generation and navigating proteome maps using mass spectrometry. Nat Rev Mol Cell Biol 11(11):789–801

Altmuller J, Palmer LJ, Fischer G, Scherh B, Wjs M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69(5):936–950

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PI, Deloukas P, Gabriel SB, Wiiliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Muzny DM, Barnes C, Darvishi K, Harles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, N mesh J, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Gonzaga-Jauregui C, Anttila V, Broder W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Zhang Q, Ghorj MI, McGinnis R, McLaren W, Takeuchi F, Grossman SR, Shlyakhert I, Hostetter EB, Sabeti PC, A dambamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrer AR, Jager SA, Chan ET, Metzger G, Veneden A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML (2009) Diversity and complexity in DNA recognition by transcription factors. Science 324(5935):1720–1723

Bartels SJ, Spruijt CG, Brinkman AB, Janssen PW, Vermeulen M, Stunnenberg HG (2011) A SILAC-based screen for methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein. PLoS ONE 6(10):e25584

Bell J, Pai A, Pickrell J, Gaffney D, Pique-Regi R, Degner J, Gilad Y, Pritchard J (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol 12(1)

Boersema P, Raimakers R, Lemeer S, Mohammed S, Heck A (2009) Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat Protoc 4(4):484–494

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33(Suppl):228–237

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132(2):311–322

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, Cherry JM, Snyder M (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22(9):1790–1797

Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436(7051):701–703

Butter F, Kappei D, Buchholz F, Vermeulen M, Mann M (2010) A domesticated transposon mediates the effects of a single-nucleotide polymorphism responsible for enhanced muscle growth. EMBO Rep 11(4):305–311

Butter F, Davison L,viturawong T, Scheibe M, Vermeulen M, Todd JA, Mann M (2012) Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. PLoS Genet 8(9):e1002982

Byrum SD, Raman A, Taverna SD, Tackett AJ (2012) ChAP-MS: a method for identification of proteins and histone posttranslational modifications at a single genomic locus. Cell Rep 2(1):198–205

Califano A, Butte AJ, Friend S, Ideker T, Schadt E (2012) Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet 44(8):841–847

Chadwick LH (2012) The NIH Roadmap epigenomics Program data resource. Epigenomics 4(3):317–324

Choudhary C, Mann M (2010) Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol 11(6):427–439

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10(3):184–194

Hum Genet (2014) 133:689–700

697

Springer
Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Sasisekhar T, Daly MJ, Wolfsberg TG, Collins FS (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 16(1):123–131

Cramer T, Cremer C (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet 2(4):292–301

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AE, Samarajiva S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinsey S, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowitz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486(7403):346–352

Cvejic A, Haer-Wigman L, Casey G, De Biasi M, Carlson C, Duggan D, James M, Liu P, whyte E, de Laat W (2012) A decade of 3C technologies: insights into nuclear organization. Genes Dev 26(1):11–24

De Godoy L, Olsen J, Cox J, Nielsen M, Hubner N, Fröhlich F, Walther T, Mann M (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455(7217):1251–1254

de Wit E, de Laat W (2012) A decade of 3C technologies: insights into nuclear organization. Genes Dev 26(1):11–24

Dejardin J, Kingston RE (2009) Purification of proteins associated with specific genomic loci. Cell 136(1):175–186

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Djeardin J, Kingston RE (2009) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65

Gibbs J, van der Brug M, Hernandez D, Traynor B, Nalls M, Lai S-L, Arepalli S, Dillman A, Rafferty I, Troncoso J, Johnson R, Ziehe H, Ferrucci L, Longo D, Cookson M, Singleton A (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet 6(5)

Gibson G (2011) Rare and common variants: twenty arguments. Nat Rev Genet 13(2):135–145

Gibson PR, Shepherd SJ (2012) Food choice as a key management strategy for functional gastrointestinal symptoms. Am J Gastroenterol 107(5):657–666; quiz 667

Glatter T, Wepf A, Aebersold R, Gstaiger M (2009) An integrated workflow for charting the human interaction proteome: insights into the PP2a system. Mol Syst Biol 5

Grossmann S, Andersson K, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park D, Griessemer D, Karlsson E, Wong S, Cabili M, Adebola R, Bamezai R, Hill A, Vannberg F, Rinn J, Genomes P, Lander E, Schaffner S, Sabeti P (2013) Identifying recent adaptations in large-scale genomic data. Cell 152(4):703–713

Gygi S, Rist B, Gerber S, Turecak F, Gelb M, Aebersold R (1999a) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17(10):994–999

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999b) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19(3):1720–1730

Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintz-Bourget J, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B (2009) Histone modifications at regulatory elements reflect global cell-type-specific gene expression. Nature 459(7243):108–112

Hesselberth J, Chen X, Zhang Z, Sabo P, Sandstrom R, Reynolds A, Thurman R, Neph S, Kuehn M, Noble W, Fields S, Stamatoyannopoulos J (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 6(4):283–289

Fried M, Crothers DM (1981) Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic Acids Res 9(23):6505–6525

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EG, Huang PY, Welbore WJ, Han Y, Ooi HS, Aariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KD, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RK, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Raboul V, Sung WK, Liu ET, Wei CL, Cheung E, Ruan Y (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462(7269):58–64

Garner MM, Revzin A (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res 9(13):3047–3060

Genomes Project Consortium, Abecasis G, Auton A, Brooks L, DePristo M, Durbin R, Handsaker R, Kang H, Marth G, McVean G (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65

Haffari G, Bashashati A, Russell R, Rafferty I, Troncoso J, Johnson R, Ziehe H, Ferrucci L, Longo D, Cookson M, Singleton A (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet 6(5)

Gibson G (2011) Rare and common variants: twenty arguments. Nat Rev Genet 13(2):135–145

Gibson PR, Shepherd SJ (2012) Food choice as a key management strategy for functional gastrointestinal symptoms. Am J Gastroenterol 107(5):657–666; quiz 667

Glatter T, Wepf A, Aebersold R, Gstaiger M (2009) An integrated workflow for charting the human interaction proteome: insights into the PP2a system. Mol Syst Biol 5

Grossmann S, Andersson K, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park D, Griessemer D, Karlsson E, Wong S, Cabili M, Adebola R, Bamezai R, Hill A, Vannberg F, Rinn J, Genomes P, Lander E, Schaffner S, Sabeti P (2013) Identifying recent adaptations in large-scale genomic data. Cell 152(4):703–713

Gygi S, Rist B, Gerber S, Turecek F, Gelb M, Aebersold R (1999a) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17(10):994–999

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999b) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19(3):1720–1730

Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintz-Norman K, Ren B, Fu XD, Topol EL, Rosenfeld MG, Frazer KA (2011) Hap DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature 470(7333):264–268

Hebert AS, Merrill AE, Bailey DJ, Still AJ, Westphall MS, Strierteri ER, Pagliarini DJ, Reeder AL, Coon JJ (2013) Neutron-encoded mass signature for之人plexed proteome quantification. Nat Methods 10(4):332–334

Heintzman ND, Hesselberth J, Chen X, Zhang Z, Sabo P, Sandstrom R, Reynolds A, Thurman R, Neph S, Kuehn M, Noble W, Fields S, Stamatoyannopoulos J (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 6(4):283–289

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis G (2012) Fast and accurate genotype imputation in
Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen AR, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shaffer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield DJ, Goble CA, Helmer-Citterich M, Eo Y, Nisbett J, Schmitt S, Curwen V, Alva AA, Guigo R, Hardison RC, Ibarra S, Harte SE, Razin A, Antonacci F, Walter JG, Badeau-Birot S, Lajoie BR, Iavarone A, Abril JF, Kellis M (2012) Systematic localization of common disease-associated variation in regulatory DNA. Nat. Genet. 37(6):699–700.
