

MEETING REPORT

Epigenomic insights into common disease

Kate Revill*¹ and Benjamin Tycko*^{2,3,4}

Abstract

A report on the Wellcome Trust Scientific Conference 'Epigenomics of Common Diseases', Hinxton, Cambridge, UK, September 13-16, 2011.

The spectacular increase in our knowledge of human genetics starting with the public and private human genome projects and coming to fruition with genome-wide association studies (GWASs) and Nextgen sequencing is impressive, but the field of human genetics has reached an impasse in which most common diseases are not fully explained by genetic variation. Epigeneticists have proposed that this gap may be filled by studying the epigenetic landscape. The recent 'Epigenomics of Common Diseases' meeting hosted by the Wellcome Trust brought together leading scientists to discuss progress in this fast-moving field. This report will highlight some of the latest developments - space limits unfortunately do not allow us to cover every talk from this comprehensive meeting so we have placed particular emphasis on genome-wide approaches that are revolutionizing epigenomics.

Progress in mapping human epigenomes

Epigenomic marks discussed included CpG methylation in DNA, histone modifications in chromatin, short-range and long-range chromatin structure, and the role of non-coding RNAs (ncRNAs). Andy Feinberg (Johns Hopkins University, USA) used microarrays and Nextgen bisulfite sequencing, with bioinformatics from Rafael Irizarry (Johns Hopkins University, USA), and found that most of the variation in methylation is not in CpG islands (CGIs), but in the 5' and 3' 'shores' of these islands. Using hematopoietic lineages, Feinberg reported that the relationship between gene expression and methylation is strongest in these regions. Based on recent data from his

laboratory he proposed that the factors leading to hyper-variability of DNA methylation in cancer may also contribute to normal tissue development and cellular identity.

Henk Stunnenberg (Radboud University, The Netherlands) covered his group's progress towards producing a blueprint of hematopoietic epigenomes. This work is being done in collaboration with Stephan Beck (University College London, UK) as a major component of the International Human Epigenome Consortium (IHEC). They are studying all major blood cell types and various leukemias using high-throughput platforms, mostly based on next-generation sequencing, to characterize genomes, methylomes and transcriptomes, as well as histone modifications. It is expected that the complete epigenomic description of well-defined easily purified human cell types will be a useful tool for understanding common diseases.

Overall, 250 distinct cell types are in large-scale epigenome mapping pipelines, including the National Institutes of Health (NIH)-funded TCGA (The Cancer Genome Atlas) project, IHEC, the NIH Epigenome Roadmap, and the ENCODE (Encyclopedia Of DNA Elements) Project. Sue Clark (Garvin Medical Institute, Australia) gave a bird's eye view talk in which she pointed out that the IHEC and ENCODE projects have overlapping and unique goals; both are mapping epigenetic marks but they tend to be using different cell types. Clark reminded the audience that certain technical aspects need to be continually revisited. Which epigenetic features should be profiled? Which assays should be used? Which cell types should be examined? An optimistic assumption is that whole blood and purified blood lineages will be useful for investigating epigenomic links to a wide array of diseases, but Clark emphasized that it is far from certain that blood cells can be used as a 'surrogate tissue' for studying diseases that have non-hematopoietic target organs. Getting large numbers of purified cell types from other lineages, such as brain and kidney, is still a problem that needs to be solved.

In addition to methylation profiling and the histone code, progress in mapping the genome for loci producing ncRNAs was also discussed. John Mattick (University of Queensland, Australia) has used deep RNA sequencing to identify 6,000 novel RNA transcripts. The majority of these sequences are dynamically transcribed, mainly into

*Correspondence: revill@cshl.edu; bt12@columbia.edu

¹Cold Spring Harbor Laboratory, New York, NY 11794, USA

²Institute for Cancer Genetics, Columbia University, 1130 St Nicholas Ave, New York, NY 10032, USA

Full list of author information is available at the end of the article

long ncRNAs (lncRNAs), of which hundreds or thousands show cell-type-specific differential expression. Functional studies in mice are now starting to define lncRNAs as major contributors to phenotypes, so this class of transcripts may well turn out to affect individual susceptibility to some common human diseases.

Bioinformatics in epigenomic mapping

Bioinformatics is a key component in sorting out the complexity of epigenetic marks. Manolis Kellis (Massachusetts Institute of Technology, USA) emphasized that the number of possible combinations of histone modifications is astronomical. To simplify and extract useful information, his group looked for combinations of histone marks that are highly recurrent at multiple genomic locations and that might track with particular functions in gene regulation. These functionally significant combinations of histone marks are now incorporated into the human genome browser at the University of California, Santa Cruz, USA. A key advance has been the ability to predict *in silico* which upstream and downstream enhancer elements are functionally relevant to which nearby genes. There are also obvious applications in interpreting non-coding variants found in GWASs.

Irizarry cited examples of misleading batch effects that affected the conclusions of profiling studies and GWASs and even led to retractions of publications. He emphasized that an initial screening of datasets by principal component analysis is good for catching such artifacts. Irizarry went on to coin the phrase 'bump hunting' to refer to picking out meaningful patterns in a noisy background of CpG methylation data. 'Smoothing' of the data involves drawing best-fit lines through the choppy data to find edges of CGIs and other key features of the epigenetic landscape. Importantly, he showed that this procedure facilitates informative whole genome bisulfite sequencing at a lower depth and thus lower total cost.

Regulation of DNA methylation and chromatin domains

In his keynote address, Peter Jones (University of Southern California, USA) presented his work on genome-wide nucleosome mapping. By using a bacterial methyltransferase to add methyl groups to accessible, nucleosome-free, promoter regions at GpC (not CpG) dinucleotides in a single bisulfite sequencing experiment, native CpG methylation and nucleosome mapping can simultaneously be performed on the same DNA molecule. This work revealed that DNA methylation encroaches over time on promoters that are wound around nucleosomes, while it remains excluded from active, nucleosome-free, promoters. This encroachment on nucleosome-occupied promoter sequences can be

explained by the known high affinity of *de novo* methyltransferases for nucleosome-bound DNA. Continuing with a related theme, Adrian Bird (University of Edinburgh, UK) described the presence of approximately 6,000 CGIs that have no association with any obvious genes or are intergenic. He terms these 'orphan islands'. Intergenic orphan islands are much more variable for methylation during development and therefore are interesting for understanding the biological function of methylation.

Recently the presence of the 'sixth base', 5-hydroxymethylcytosine (5hmC), has provided a possible explanation for rapid active or passive cytosine demethylation. Wolf Reik (University of Cambridge, UK) discussed mechanisms of epigenetic reprogramming focusing on TET1, an mC-hydroxylating enzyme that is produced from a gene that becomes methylated and silenced with cell differentiation. His experiments suggest that production of TET1 in embryonic cells is indeed important for the cytosine demethylation that is a key feature of epigenetic reprogramming early in development. Amanda Fisher (Imperial College London, UK) finds increased 5hmC at several silent genes in human B cell-mouse embryonic stem cell heterokaryons immediately after reprogramming. This reprogramming is blocked in embryonic stem cells deficient for PC1 or PC2, indicating that the polycomb complexes are crucial. In a presentation on the classic epigenetic model system X chromosome inactivation, Edith Heard (Institute Curie, France) showed evidence for novel regulatory elements within the highly complex and multipartite 10 Mb XIC region. She finds that this region is riddled with regulatory transcription factor sites and chromatin immunoprecipitation-sequencing (ChIP-SEQ) peaks, suggesting that we need to look beyond just chromatin and examine long-range subnuclear organization. Her allele-specific chromosome conformation capture data are revealing megabase-scale DNA domains with a preponderance of specific chromatin markings. This same concept has also emerged from recent DNA methylation mapping studies in cancer cells by the Clark and Feinberg laboratories, which led to some repartee on nomenclature ('blobs?') for this novel and possibly fundamental type of long-range epigenomic structure.

Genetic-epigenetic interactions

Not unexpectedly, given the strong interest of human geneticists in developing 'post-GWAS' approaches for studying complex diseases, genetic-epigenetic interactions turned out to be one of the recurrent themes. We (Tycko) discussed published and unpublished data from our group showing that genetic haplotypes can exert a major influence on DNA methylation patterns. For many loci, the methylation status of CpG dinucleotides can be

predicted simply by knowing the genotype at adjacent SNPs. This *cis*-acting genetic-epigenetic relationship, haplotype-dependent allele-specific methylation (ASM), was first revealed using SNP arrays, and it is now being studied using Nextgen bisulfite sequencing. Mapping ASM across human epigenomes has a practical application, namely to find regulatory SNPs and haplotypes, which betray their presence by conferring a physical asymmetry in DNA methylation between the two alleles. Some of these SNPs and haplotypes will co-map with GWAS peaks, and this genetic-epigenetic co-mapping can provide molecular proof that the GWAS signal is a true positive, reflecting the presence of a *bona fide* regulatory variant.

This idea is being developed by other laboratories, including Jon Mill's research group (King's College London, UK). His laboratory has focused on DNA methylation profiling of brain tissue, towards an understanding of the epigenetics of neuropsychiatric and neurodegenerative diseases. Using methylation-dependent immunoprecipitation he can distinguish among brain cortical regions by their methylation signatures. Slightly different from Feinberg's findings in other tissues, intragenic CGIs and non-CGI promoters are the most abundant tissue-specific differentially methylated regions in Mill's data from brain.

Epigenome-wide association studies: are EWASs the new GWASs?

Applying ASM mapping and related modalities such as allele-specific expression and DNase hypersensitivity mapping to extract maximum information from GWASs is straightforward, but making links to common diseases using epigenetic information by itself will be much more challenging. Just as this took some time in GWASs, the ground rules for study design and statistical analysis in the epigenome-wide association study (EWAS) field are only just starting to emerge. To control for shared genetic factors twin studies will be very important: while methylation profiling in monozygotic and dizygotic twins reveals that methylation patterns differ between twin pairs but are mostly similar within pairs (reflecting the *cis*-acting influence of shared haplotypes on DNA methylation patterns), research in monozygotic twins with high discordance rates for common diseases suggests that environmental or stochastic epigenetic factors can also produce some differences within pairs. This concept is at the core of the EWAS approach, and it has motivated several large cohort studies using twin pairs. Talks by Vardhman Rakyan (Queen Mary University of London, UK), Tim Spector (Kings College London, UK) and Stephen Kingsmore (National Center for Genome Resources, USA) presented some very early findings from these types of longitudinal studies.

Environmental influences on epigenomes

Inter-individual epigenetic variability can occur at any point in an individual's lifetime but *in utero* development is a key period during which the epigenome is susceptible to environmental exposures such as infection, poor diet or other types of maternal stress. In a thoughtful presentation on epigenetics and the determination of phenotypes Emma Whitelaw (University of Western Australia, Australia) emphasized that the apparently simple sequence of (exposure → change in DNA methylation → phenotype) may not really be so simple; it could alternatively be (exposure → altered cell types → phenotype), in which case the change in methylation is essentially an artifact of the altered cellular composition. So we need to be careful in interpreting data as to whether epigenetic marks are 'instructive' or causal, versus secondary. She suggested that it may well be that only special types of promoters are influenced by environmental effects on CpG methylation, for example, the mouse *A^y* allele, which is a retroviral long terminal repeat insertion in the *Agouti* locus that confers sensitivity of coat color to maternal diets. While this special case has raised the idea that alleles like this may exist in humans, in fact it is difficult to find data that confirm this idea without a possible artifactual explanation. With regard to possible transgenerational epigenetic effects, it is important to remember that such effects do not necessarily reflect true gametic transmission of an epigenetic mark; uterine environment, maternal health and infectious agents are other possibilities. This caveat was emphasized by Oliver Rando (University of Massachusetts, USA) whose experiments in mice are designed to test for paternal not maternal effects; to exclude artifacts of the uterine environment as much as possible. They asked whether paternal diet influences gene expression in the offspring and found a group of differentially expressed genes. However, it is not yet known whether this effect on gene expression is due to altered epigenetics in sperm: there could be alternative explanations.

Future perspectives

From this conference it was clear that mechanistic and mapping studies in the new field of epigenomics are making great strides. However, the question of how to best utilize epigenomic data for uncovering novel disease loci remained very much open. An important panel discussion outlined the challenges in interpreting epigenetic profiles, which by nature are phenotypic, tissue specific and dynamic and thus prone to confounders and so-called 'reverse causation'. While the idea of using epigenetic mapping as a tool in conjunction with standard GWASs was well accepted, it was emphasized that the pure EWAS approach cannot, by itself,

distinguish the direction of the relationship between disease and epigenetic variation. There was no easy answer, so it will be important to revisit this question after the initial EWAS results are analyzed and vetted for reproducibility.

Abbreviations

ASM, allele-specific DNA methylation; CGI, CpG island; ENCODE, Encyclopedia of DNA Elements; EWAS, epigenome-wide association study; GWAS, genome-wide association study; 5hmC, 5-hydroxymethylcytosine; IHEC, International Human Epigenome Consortium; lncRNA, long non-coding RNA; ncRNA, non-coding RNA; NIH, National Institutes of Health; SNP, single nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by grants R01AG036040-01, R01AG035020-01, P01-HD035897, R01 MH092580-01A1 and DP3DK094400-01, all from the National Institutes of Health.

Author details

¹Cold Spring Harbor Laboratory, New York, NY 11794, USA. ²Institute for Cancer Genetics, Columbia University, 1130 St Nicholas Ave, New York, NY 10032, USA. ³Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, 1130 St Nicholas Ave, New York, NY 10032, USA. ⁴Department of Pathology, Columbia University, 1130 St Nicholas Ave, New York, NY 10032, USA.

Published: 15 November 2011

doi:10.1186/gm287

Cite this article as: Revill K, Tycko B: Epigenomic insights into common disease. *Genome Medicine* 2011, **3**:71.