Epigenetic biomarker development

Christoph Bock

Max-Planck-Institut für Informatik, Saarbrücken, Germany
Correspondence: e-mail: cbock@mpi-inf.mpg.de, phone: +49 681 9325 303, fax: +49 681 9325 399.

Keywords:
Molecular diagnostics, DNA methylation, epigenome, computational epigenetics, cancer, complex diseases, diagnostics, therapy optimization, personalized medicine, genome-wide association studies

Executive summary:
Epigenetic regulation affects human disease

- Epigenetic alterations have a well-established and causal role in cancer.
- Mechanisms of epigenetic regulation appear to be involved in autoimmune diseases, neural disorders and other complex diseases.

Epigenetic biomarkers inform treatment decisions

- Epigenetic biomarkers measure disease-associated and drug-associated epigenetic alterations, thus providing decision support for routine clinical treatment and drug discovery.
- Practical advantages of epigenetic biomarkers include low susceptibility to short-term fluctuations and straightforward sample processing (DNA methylation analysis works well on paraffin samples).

A systematic approach to epigenetic biomarker development

- Biomarker development projects should adopt a systematic approach to genome-scale screening, computational optimization and high-throughput validation.

Relevant bioinformatic tools

- Consistent use of bioinformatic tools for data processing, candidate prioritization and assay optimization can significantly increase the speed and success rate of biomarker development.

Open questions

- To increase the efficiency of epigenetic biomarker development, the following topics need to be better understood: (i) genome-scale search vs. candidate approach, (ii) sample size and statistical power, (iii) success criteria, and (iv) combining genetic/epigenetic biomarkers.

Conclusion

- Epigenetic biomarkers have not yet realized their potential for informing clinical decisions.
- An integrative approach to epigenetic biomarker development may overcome current roadblocks.

Future perspective

- In the coming years, epigenome-wide association studies will increasingly complement genome-wide association studies in the search for novel disease genes and clinically relevant biomarkers.

|

Summary

Epigenetic mechanisms control gene expression in a way that is stably propagated over multiple cell divisions, but which is also flexible enough to respond to environmental influences. This intermediate position between stability and plasticity renders epigenetic information highly useful for monitoring cellular states in the context of personalized medicine. Epigenetic alterations have also been identified as causal events for common diseases

such cancer and autoimmune disorders. The goal of epigenetic biomarker development is to design experimental assays that produce relevant information for diagnosis, prognosis and therapy optimization in routine clinical treatment and drug discovery. Here, we outline a systematic approach to epigenetic biomarker development and highlight key bioinformatic tools that facilitate discovery, optimization and validation of novel biomarkers.

Epigenetic regulation affects human disease

Epigenetic mechanisms such as DNA methylation and histone modifications regulate gene expression by modulating the packaging of the DNA inside the nucleus [1, 2]. Epigenetic information is faithfully propagated over multiple cell divisions in somatic cells. However, it is altered during cellular differentiation and largely erased in the germline and the early embryo [3]. Furthermore, environmental and lifestyle-related influences such as nutrition and exposure to stress can induce epigenetic alterations [4, 5]. In this sense, the human epigenome can be regarded as a biochemical record of relevant life events, accumulating alterations over a lifetime. Consistent with this view, it has been shown that monozygotic twins start off with highly similar epigenomes that diverge with age, at a rate that is decreased when twins share a common environment [6].

While many epigenetic alterations may constitute epigenetic drift without phenotypic effect, it is evident that some environment-induced changes modulate gene expression in disease-relevant ways [7-9]. It is thus little surprising that epigenetic deregulation has been linked to a variety of common diseases. For example, epigenetic mechanisms regulate autoimmunity, and their deregulation has been shown to contribute to rheumatoid arthritis and systemic lupus erythematosus [10]. Similarly, neural activity in the brain is epigenetically regulated [11], and it has been extensively speculated about the relevance of epigenetic alterations for mental disorders such as schizophrenia [12] and substance abuse [13]. Conclusive and well-replicated results are still missing for most diseases, but the data collected so far strongly support the prospect that adequately powered epigenome-wide association studies will identify key epigenetic modulators in neural disorders, metabolic diseases and non-disease phenotypes such as body height and weight.

Of all common diseases, the role of epigenetic alterations in cancer has been analyzed in highest detail [14, 15]. It is now clear that epigenetic silencing of tumor suppressor genes is a frequent event in multiple cancers, and several lines of evidence support its causal involvement in cancer progression [16, 17]. Furthermore, epigenetic similarities between cancer cells and adult stem cells suggest that epigenetic deregulation may program cells for a cancer fate behavior long before they are visually identifiable as tumor cells [18].

Epigenetic biomarkers inform treatment decisions

The goal of clinical biomarkers is to provide physicians with relevant information about the presence or absence of a disease (diagnostic biomarkers) as well as about patient and disease characteristics that influence treatment decisions (prognostic and therapy-optimization biomarkers). Conceptually, epigenetic biomarkers consist of two complementary building blocks: (i) an experimental assay that provides accurate measurements of epigenetic alterations in a given patient sample, either at a single locus or at multiple genomic regions; and (ii) a sample classifier that translates the experimental read-out into the biomarker outcome, e.g. the predicted disease subtype or tumor grading.

Many mechanisms of epigenetic regulation have been discovered in recent years, DNA methylation, histone methylation, histone acetylation, microRNAs and other non-coding RNA being among the most prominent [19, 20]. However, epigenetic biomarker development has so far focused mostly on DNA methylation (Figure 1), both because of practical considerations (DNA methylation is relatively stable and easy to measure with current technologies [16, 21]) and because of the well-established role of DNA methylation in cancer [14, 15]. Reflecting its prevalence in current applications, the remainder of this paper concentrates on DNA methylation biomarkers; though we note that epigenetic biomarkers based on histone modifications [22] and non-coding RNAs [23, 24] may increasingly complement DNA methylation biomarkers in the near future.

The frequency with which DNA methylation patterns are altered in early-stage tumors [25, 26] has fueled efforts to develop diagnostic biomarkers, utilizing cancer-specific alterations such as hypermethylation of promoter regions for early detection of potentially fatal tumors [27]. To be suitable for population screening, diagnostic biomarkers are usually based on readily available body fluids. For example, stool DNA has been used for detecting colon cancer [28, 29], urine for prostate cancer [30, 31] as well as bladder cancer [32, 33], and tampons / Pap smears for endometrial as well cervical cancer [34, 35]. Furthermore, extensive efforts have been made to develop accurate biomarkers based on blood, not only for leukemias (where blood is a directly affected tissue [36, 37]), but also for a number of solid tumors.

The rationale behind blood tests for detecting solid tumors is that tumor cells may shed epigenetically altered DNA into the bloodstream [27] or that blood cells may have undergone epigenetic changes representative of those present in the tumor, e.g. in response to specific environmental influences [38]. Both concepts appear problematic: the amount of circulating tumor DNA is rarely sufficient for detection against the epigenetic heterogeneity of normal blood [39], and it is doubtful whether blood is a suitable proxy for epigenetic alterations elsewhere. Nevertheless, there have been successes in developing blood-based diagnostic biomarkers. One of the most extensively validated examples is the *septin 9* gene, which exhibits a significant degree of hypermethylation in the blood plasma of more than 50% of colon cancer patients, but is unmethylated in more than 90% of healthy patients [40-42]. The clinical value of such moderately sensitive and specific biomarkers is an open question, depending highly on the invasiveness of follow-up (high invasiveness renders false positives costly) and the treatment options for improving cancer survival (low treatment success rates reduce the value of true positives).

Complementary to early diagnosis of developing tumors, a second class of epigenetic biomarkers aims to support clinical decision making once a tumor has been identified. Such prognostic and therapy-optimization biomarkers help address the following questions by measuring relevant aspects of tumor biology: (i) Is the pathological diagnosis confirmed on the molecular level? (ii) Does the tumor fall into a known subclass of epigenetically characterized tumors (such as CpG island methylator phenotype [43])? (iii) Can and should the tumor be treated? (iv) Which type of therapy is appropriate? (v) Which doses should be used for radiation and chemotherapy? (vi) How strong will the side effects be and how can they be minimized?

Epigenetic biomarkers for prognosis and therapy optimization address a conceptually easier problem than diagnostic biomarkers, for two reasons. First, biopsy material of the primary disease tissue is usually available once the diagnosis has been confirmed, while diagnostic biomarkers have to rely on more readily available tissues (e.g. blood), which may not represent the disease tissue well. Second, specificity requirements are orders of magnitude lower for prognostic and therapy-optimization biomarkers than for diagnostic biomarkers. This is because diagnostic biomarkers are ultimately designed for screening a largely healthy population, such that relatively small false positive rates can already give rise to a substantial number of unnecessary follow-up examinations in healthy individuals. Hence, it is hardly surprising that several of the more compelling examples of epigenetic biomarkers focus on identifying disease subtypes with distinct clinical properties [26, 44], rather than distinguishing between diseased patients and healthy controls. For example, Hegi et al. showed that aberrant methylation in the promoter region of the *MGMT* gene is a highly significant predictor of chemotherapy resistance to alkylating agents in glioblastomas. In their study, temozolomide treatment resulted in a median survival benefit of six months for the *MGMT*-methylated but not for the *MGMT*-unmethylated patients. Building upon this milestone paper, we and others have recently developed optimized biomarkers for routine clinical testing of *MGMT* promoter methylation [45, 46], such that oncologists can increasingly take the *MGMT* methylation status into account when devising individualized therapy regimes for glioblastoma patients.

A systematic approach to epigenetic biomarker development

Recent advances in high-throughput sequencing [47, 48] for the first time enable epigenetic biomarker discovery on a truly genome-wide scale. This increased coverage is likely to uncover many new genomic regions that exhibit disease-specific epigenetic alterations, including those that are located outside well-known candidate regions such as CpG islands and gene promoters [25]. However, the steep increase in the scale of the search gives rise to significant bioinformatic challenges, such as channeling high-throughput sequencing data through an effective data-reduction and target-identification pipeline, and addressing the statistical challenges of a massive multiple-testing problem. Furthermore, increased efforts are required to ensure rapid translation of candidate biomarkers into diagnostic tools that address clinicians' information requirements, rather than stockpiling ever larger numbers of unreplicated differentially methylated regions.

We believe that a systematic approach to biomarker development can overcome many of these challenges. In Figure 2, we outline a procedure that we personally regard as useful for epigenetic biomarker development. This procedure is based on three key concepts: (i) to maximize genomic coverage in the early stages of the search (steps 1 and 2); (ii) to employ computational methods for identifying and optimizing a small number of highly promising candidate biomarkers (steps 3 and 4); and (iii) to validate biomarker performance in large cohorts using highly targeted assays (steps 5 and 6). The proposed procedure can be applied to any well-defined disease state for which at least two high-quality case-control cohorts are available (the primary cohort being used for biomarker development and the validation cohort for an unbiased assessment of biomarker performance). Typical application scenarios include the search for a diagnostic biomarker of early-stage colon cancer (based on stool DNA) or the identification of a therapy-optimization biomarker predicting optimal azacitidine doses in myelodysplastic syndrome patients.

Because genome-scale analysis of DNA methylation is still a costly exercise, large-scale epigenetic biomarker development becomes more feasible when several experimental methods with very different trade-offs between genomic coverage and per-sample costs are combined. For the initial discovery phase (step 1), we advocate the use of experimental methods that provide maximum genomic coverage [49-52], even when high per-sample costs severely restrict the number of samples that can be processed in this phase. Subsequently, the uncertainty arising from small sample size in the initial screening phase is addressed by a medium-scale confirmation phase, in which a liberal selection of candidate regions from the initial screening are evaluated in a fivefold larger number of samples. Experimental methods for medium-scale confirmation must be highly customizable and able to assess the DNA methylation status of thousands of CpGs in up to a hundred samples. These requirements are met by bisulfite sequencing of region-specifically enriched DNA [53, 54] and by epigenotyping assays such as Illumina Infinium HumanMethylation27 (http://www.illumina.com/pages.ilmn?ID=243). Based on sensitivity and specificity estimates derived from the results of the medium-scale confirmation phase, a small number of highly predictive genomic regions are selected as candidate biomarkers. For each of these epigenetically altered regions, an optimized assay is developed that tests for DNA methylation at a small number of representative CpGs. This step exploits the fact that DNA methylation status is highly correlated between adjacent CpG dinucleotides [55], such that assaying a handful of carefully selected CpGs often provides an accurate DNA methylation read-out of an entire CpG island or gene promoter [46, 56]. Several experimental methods enable robust DNA methylation measurement of a small number of CpGs at low cost, addressing key requirements of biomarker validation and subsequent clinical use. Bisulfite pyrosequencing [57], Ms-SNuPE [58], COBRA [59] and mass spectrometry [60] provide quantitative DNA methylation information for individual CpGs, conferring increased robustness against random fluctuations. In contrast, MethyLight [61] and MSP [62] query the DNA methylation status of several CpGs simultaneously, enabling highly sensitive detection of specific methylation patterns. Finally, clonal bisulfite sequencing – which is commonly regarded as the gold standard for DNA methylation analysis [63] – is useful for assay quality control and for identifying the most representative CpGs within a given region (P. Schüffler, T. Mikeska, T. Lengauer and C. Bock, submitted), but is too laborious for routine clinical use.

Relevant bioinformatic tools

Essentially all steps outlined in Figure 2 utilize bioinformatic tools, for tasks ranging from sample selection over candidate ranking, assay design and biomarker optimization to the final exercise of unbiased performance evaluation. We believe that intelligent use of computational methods can help overcome many of the obstacles that threaten to delay biomarker development or diminish the value of the end product. To give a concrete example, automatic cross-checking against SNP databases and the computational simulation of unknown SNPs (as implemented in MethMarker, http://methmarker.mpi-inf.mpg.de/) facilitate selection of candidate biomarkers that are robust toward DNA sequence variation. Similarly, computational prediction of epigenetic variation helps focus on candidate biomarkers that exhibit little variation among healthy individuals [64], which minimizes the risk of false classifications due to population heterogeneity. In the following, we provide a brief outline of the bioinformatic tools that we find particularly useful for epigenetic biomarker development, and we sketch how these tools contribute to the different steps of the procedure depicted in Figure 2. A discussion of additional bioinformatic methods and software tools for epigenetic research is available from our recent review on computational epigenetics [65].

The first bioinformatic challenge of epigenetic biomarker development is to select the highest-quality and most representative patient samples for genome-scale screening (step 1 in Figure 2). While this choice is sometimes dictated by practical considerations (such as the amount of available DNA), it is often possible to use existing data to select a subset of cases from a larger cohort. For example, samples can be clustered by their gene expression profiles, in order to make sure that all relevant disease subtypes are covered, or to focus on just a single disease subtype. Suitable software packages for performing these kinds of analysis include GenePattern [66] and R/Bioconductor [67]. Once DNA methylation mapping has been performed on the selected samples, bioinformatic tools become critical for data processing. The necessary steps vary between different experimental methods, but usually include read alignment [68, 69], data normalization [50, 70] and visualization as well as data analysis [71, 72].

In the medium-scale confirmation phase (step 2 in Figure 2), a genome processing tool such as Galaxy [73] helps assemble a list of genomic regions that show evidence of disease-specific alterations in the initial screening. On this basis, a statistics software such as R/Bioconductor [67] can be used to select candidate regions and to prepare the specification files necessary for ordering customized hybrid-selection probes [53, 54] or epigenotyping microarrays (www.illumina.com/downloads/InfMethylation_AppNote.pdf). Once the data from the confirmation phase have been generated, bioinformatic tools facilitate data processing and quality control, using vendor-provided software (e.g. Illumina BeadStudio) or its open-source alternatives [74, 75]. Based on the preprocessed data, preliminary sensitivity and specificity estimates can then be calculated with statistical learning software such as the Weka data mining suite [76] or the multi-purpose statistical workhorse R/Bioconductor [67].

Selecting the most promising biomarkers for validation (step 3) is arguably the most crucial and challenging step of the procedure outlined in Figure 2. Clearly, the main selection criterion is the predictive power for the disease condition of interest, which can be estimated from the experimental data accumulated in steps 1 and 2. However, to maximize the chances of selecting biomarker candidates that validate well and that may also provide new insights into disease mechanisms, further data should be taken into account. On the one hand, bioinformatic methods can be used to predict the inherent propensity with which a given genomic region is involved in normal [77, 78] and disease-specific epigenetic regulation [79-81]. The idea is that observed epigenetic alterations which contradict these predictions are likely to be the result of positive somatic selection and thus functionally involved in cancer – which makes them particularly strong biomarker candidates. On the other hand, statistical comparison with public databases such as Oncomine [82] and MethCancerDB [83] as well as the use of pathway analysis tools [84] can help identify candidate biomarkers that relate to known molecular pathways and can give rise to hypotheses about a mechanistic link between the epigenetic alteration and the disease condition. (While good biomarkers do not necessarily measure epigenetic alterations that are causal for the disease condition they predict, a plausible mechanistic model can significantly increase the credibility of a biomarker candidate.)

To enable validation in large cohorts and subsequent clinical use, a targeted assay has to be developed specifically for each candidate region (step 4), in such a way that it maintains the predictive power for the disease condition (as established in steps 1 to 3) but is significantly more cost-efficient, robust and easy-to-handle. A number of experimental methods are suitable for this purpose (see previous section), and bioinformatic methods have been published that facilitate key steps of assay design and data analysis. BiSearch [85] and Methyl Primer Express (http://www.appliedbiosystems.com/methylprimerexpress) are widely used for methylation-specific primer design; BiQ Analyzer [86] and QUMA [87] support data analysis for bisulfite sequencing; commercial packages such as PyroMark Assay Design Software (http://www.pyrosequencing.com/DynPage.aspx?id=7257) and EpiDesigner (http://www.epidesigner.com/) facilitate custom assay design for the respective methods; and the MassArray R package [88] provides an open-source alternative for EpiTYPER data analysis. However, none of these software packages is specifically designed for biomarker development, which is why we have recently developed MethMarker (http://methmarker.mpi-inf.mpg.de/, Schüffler et al. submitted) as a dedicated software tool for optimization and validation of DNA methylation biomarkers. Briefly, MethMarker implements assay design for multiple experimental methods within a single interface, it utilizes DNA methylation profiles of representative cases and controls to identify predictive CpGs, and it estimates the sensitivity and specificity of candidate biomarkers using logistic regression models.

Once the targeted DNA methylation assay has been applied to all samples from the primary case-control cohort, initial classification models are trained to distinguish between cases and controls. Depending on the number of candidate biomarkers and the difficulty of the classification problem, this classification model may be as simple as a single threshold to which the DNA methylation measurements are compared, or as complex as a logistic regression model integrating the measurements of multiple assays into a single class prediction. Using statistical learning software and cross-validation on the training data [89, 90], the most suitable classification model is selected and a preliminary performance assessment is derived. Based on these results, the most promising biomarker candidates are selected for confirmation in an independent validation cohort with similar properties as the primary case-control cohort (step 5 in Figure 2). If a biomarker validates well in retrospective analysis on several cohorts, it may become worthwhile to conduct a prospective study, which can provide the most conclusive proof of a biomarker's predictive value. On the other hand, if multiple candidate biomarkers fail to replicate during the validation phase, it is critical to identify the reason and to adjust the procedure accordingly. Typical problems include insufficient sample size during the training phase, differences in population structure or disease subtype between primary cohort vs. validation cohort, and changes in the experimental protocol (analysis performed in a different lab, samples classified by different pathologists, etc.). Because validation studies depend highly on biostatistical methods, they are best performed using statistics packages such as R (http://www.r-project.org/), SAS/STAT (http://www.sas.com/technologies/analytics/statistics/stat/index.html) or SPSS PASW (http://www.spss.com/statistics/). Furthermore, tools supporting data integration and clinical trials management [91] can facilitate the complex logistics of large-scale biomarker validation.

Open questions

Although researchers have been working on epigenetic biomarkers for more than a decade [27], some aspects relating to the design of biomarker development projects have not been conclusively addressed:

(i) What is the value of genome-wide biomarker discovery as opposed to candidate gene approaches? A sizable number of genes have been found hypermethylated in more than one cancer [16], suggesting that these genes might be good biomarker candidates for other cancers as well. Indeed, assays are now available to test the DNA methylation status of a moderate number of cancer-related genes at low cost [92, 93]. In contrast, experiences from cancer genome sequencing [94, 95] suggest that our knowledge of cancer biology is still insufficient to confidently pick candidate regions, arguing for a more unbiased genome-scale approach. The procedure outlined in Figure 2 aims to combine both approaches by starting genome-wide (step 1) but taking prior knowledge into account when selecting biomarker candidates for optimization and validation (step 2).

(ii) What sample sizes are required for epigenetic biomarker development? In order to maximize the probability that newly discovered biomarkers replicate well in independent patient cohorts, biomarker development projects should be based on a sufficiently large primary case-control cohort. Specifically, the sample size needs to be high when the difference between cases and controls is small, when only a small number of patients carry

the relevant epigenetic alteration and when many genomic regions are tested in parallel. While the question of adequate sample size has been thoroughly addressed for genome-wide association studies [96, and references therein], systematic power studies for epigenetic biomarker development are still missing. For this reason, the proposed sample sizes in Figure 2 should be regarded as rough estimates derived from the literature as well as from our own experiences, and not as definite points of reference based on robust statistical calculations.

(iii) What are suitable success criteria to guide epigenetic biomarker development? In each of the steps outlined in Figure 2, a relatively small number of candidate biomarkers are selected for further analysis, while all other genomic regions are discarded. A pragmatic approach to this selection problem is to rank all regions by their potential as candidate biomarkers and to choose a fixed number of promising candidates from the top of the ranked list. Nevertheless, further research should aim at identifying "hard" criteria for deciding which regions warrant detailed follow-up. For example, it may turn out that half of the differentially methylated regions with a raw p-value below 10^{-6} replicate well in validation cohorts, while few of those with a p-value above 10^{-4} do, providing an empirical indication of what level of statistical stringency is adequate. (Classical multi-testing correction is not easily applicable to epigenome-wide data because the epigenetic states of adjacent regions are highly correlated, such that the actual number of statistically independent tests may be substantially lower than the number of genomic regions covered.) Furthermore, the clinical uptake of validated biomarkers needs to be monitored in order to better understand which levels of sensitivity and specificity have to be achieved by diagnostic, prognostic and therapy-optimization biomarkers, respectively, to have a sizable impact on clinical practice.

(iv) What is the benefit of combining epigenetic biomarkers with other types of biomarkers? The search for disease biomarkers is by no means restricted to epigenetic alterations [97] – which gives rise to the question whether the predictiveness of epigenetic biomarkers can be improved by integration with genomic or proteomic biomarkers. Two hypotheses seem plausible. On the one hand, different types of biomarkers may just measure different but highly correlated aspects of the same overall disease state, such that their combination adds little value. This model is consistent with the observation that epigenetic alterations sometimes go hand-in-hand with genetic alterations [26, 98], in which case it may be sufficient to measure either one. On the other hand, it has been proposed that genetic and epigenetic alterations provide alternative routes for tumor cells to acquire the hallmarks of cancer [99, 100], suggesting that a biomarker monitoring genetic as well as epigenetic alterations might be substantially more accurate than an exclusively epigenetic biomarker. Current results seem to favor the second model [101-103], but more research is clearly warranted.

Conclusion

Epigenetic biomarkers hold great promise for improving cancer therapy. In many cases, aberrant methylation is detectable already in early-stage and pre-malignant tumors [14, 18, 27], when surgical treatment can be highly effective. Furthermore, specific DNA methylation patterns often correlate with clinical parameters such as cancer stage, survival time and chemotherapy resistance, which gives rise to new opportunities for informed treatment decisions and survival prognosis, thus enabling more personalized cancer therapy. However, in spite of a number of recent successes [26, 44], DNA methylation biomarkers have been slow to generate measurable impact on clinical cancer therapy. Beyond a number of conceptual reasons discussed elsewhere [97], the gap between discovery and clinical adaptation is likely aggravated by inefficiencies of the biomarker development process itself, leading to promising biomarkers being missed or discarded, while poor candidates fail at late stages of the validation process. We believe that a systematic approach to biomarker development can help overcome some of these issues, especially when it leverages the use of bioinformatic methods for data integration and decision support. Figure 2 outlines a procedure that is tailored to the development of DNA methylation biomarkers in cancer and other complex diseases. Compared to biomarker phase diagrams published elsewhere [e.g. in 104], this procedure addresses the specific requirements of epigenetic biomarker development, and it highlights software tools that can provide critical guidance for deciding which candidate biomarkers should be discarded and which should be carried through to validation in large cohorts. We are confident that systematic, bioinformatics-driven strategies can increase the efficiency and reduce the cost of biomarker development projects, thus helping to fulfill the clinical promise of epigenetic biomarkers in cancer and beyond.

Future perspective

Published research on epigenetic biomarkers has largely been confined to cancer. However, epigenetic testing becomes an increasingly attractive option for researchers working on other common diseases. An important hypothesis driving the field is that epigenetic alterations may link life events and environmental exposures to disease risk, thus providing a biochemical record of risk factor exposure that is almost impossible to reconstruct based on genetic and environmental data alone. Recent technical advances have also done their part, by making epigenome mapping increasingly cost-efficient and less cumbersome, such that it is now a valid option for many labs to screen for epigenetic alterations in their disease of interest. It is not difficult to predict that the search for epigenetic alterations and disease-specific biomarkers will continue on a genome-wide scale, swiftly following the example of genome-wide association studies (GWAS) and their search of disease-causing DNA sequence polymorphisms [96]. In fact, several theoretical frameworks have already been proposed for integrating and combining the power of genetic and epigenetic association studies on a genome-wide scale [105-107].

While it is likely that the integrated search for genetic and epigenetic risk factors will ultimately result in a much improved understanding of complex diseases, major stumbling blocks exist on the road ahead. Specifically, it appears that epigenetic alterations in complex diseases other than cancer are orders of magnitude weaker and rarer than those observed in tumors, which would render essentially all published epigenome-wide association studies for non-cancer diseases severely underpowered. Experience from the early days of GWAS suggests that the findings of underpowered studies rarely replicate and that substantial improvements in sample size and statistical rigor had to be made before the GWAS field could assume the central role in discovering disease genes that it has assumed during the last few years [108]. We believe that a focus on developing (and publishing) a small number of validated epigenetic biomarkers rather than large lists of unreplicated differentially methylated regions can help address these concerns, and we conclude by underlining the critical importance of validating epigenetic biomarkers in more than a single cohort.

Acknowledgements

We would like to thank Thomas Lengauer, Jörn Walter, Thomas Mikeska, Peter Schüffler, Yassen Assenov and the members of the CANCERDIP consortium (http://www.cancerdip.eu/) for helpful discussions. This work was partially funded by the European Union through the CANCERDIP project (HEALTH-F2-2007-200620).

Competing interests

The author declares that no competing financial interests exist.

References

1. Kouzarides, T: Chromatin modifications and their function. *Cell* 128, 693-705 (2007).
2. Weber, M,D Schübeler: Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr Opin Cell Biol* 19, 273-80 (2007).
3. Reik, W: Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447, 425-32 (2007).
4. Heijmans, BT, EW Tobi, AD Stein *et al.*: Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc Natl Acad Sci U S A* 105, 17046-9 (2008).
5. McGowan, PO, A Sasaki, AC D'Alessio *et al.*: Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nat Neurosci* 12, 342-8 (2009).

6. Fraga, MF, E Ballestar, MF Paz *et al.*: Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A* 102, 10604-9 (2005).
7. Hirst, M,MA Marra: Epigenetics and human disease. *Int J Biochem Cell Biol* 41, 136-46 (2009).
8. Feinberg, AP: Phenotypic plasticity and the epigenetics of human disease. *Nature* 447, 433-40 (2007).
9. Bjornsson, HT, MD Fallin,AP Feinberg: An integrated epigenetic and genetic approach to common human disease. *Trends Genet* 20, 350-8 (2004).
10. Richardson, B: Primer: epigenetics of autoimmunity. *Nat Clin Pract Rheumatol* 3, 521-7 (2007).
11. Mehler, MF: Epigenetics and the nervous system. *Ann Neurol* 64, 602-17 (2008).
12. Petronis, A: The origin of schizophrenia: genetic thesis, epigenetic antithesis, and resolving synthesis. *Biol Psychiatry* 55, 965-70 (2004).
13. Kalsi, G, CA Prescott, KS Kendler,BP Riley: Unraveling the molecular mechanisms of alcohol dependence. *Trends Genet* 25, 49-55 (2009).
14. Esteller, M: Epigenetics in cancer. *N Engl J Med* 358, 1148-59 (2008).
15. Feinberg, AP,B Tycko: The history of cancer epigenetics. *Nat Rev Cancer* 4, 143-53 (2004).
16. Esteller, M: Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* 8, 286-98 (2007).
17. Jones, PA,SB Baylin: The epigenomics of cancer. *Cell* 128, 683-92 (2007).
18. Feinberg, AP, R Ohlsson,S Henikoff: The epigenetic progenitor origin of human cancer. *Nat Rev Genet* 7, 21-33 (2006).
19. Berger, SL, T Kouzarides, R Shiekhattar,A Shilatifard: An operational definition of epigenetics. *Genes Dev* 23, 781-3 (2009).
20. Bird, A: Perceptions of epigenetics. *Nature* 447, 396-8 (2007).
21. Suzuki, MM,A Bird: DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* 9, 465-76 (2008).
22. Kondo, Y, L Shen, AS Cheng *et al.*: Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nat Genet* 40, 741-50 (2008).
23. He, L, JM Thomson, MT Hemann *et al.*: A microRNA polycistron as a potential human oncogene. *Nature* 435, 828-33 (2005).
24. Lu, J, G Getz, EA Miska *et al.*: MicroRNA expression profiles classify human cancers. *Nature* 435, 834-8 (2005).
25. Irizarry, RA, C Ladd-Acosta, B Wen *et al.*: The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet* 41, 178-86 (2009).
26. Weisenberger, DJ, KD Siegmund, M Campan *et al.*: CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat Genet* 38, 787-93 (2006).
27. Laird, PW: The power and the promise of DNA methylation markers. *Nat Rev Cancer* 3, 253-66 (2003).
28. Glockner, SC, M Dhir, JM Yi *et al.*: Methylation of TFPI2 in stool DNA: a potential novel biomarker for the detection of colorectal cancer. *Cancer Res* 69, 4691-9 (2009).
29. Müller, HM, M Oberwalder, H Fiegl *et al.*: Methylation changes in faecal DNA: a marker for colorectal cancer screening? *Lancet* 363, 1283-5 (2004).
30. Payne, SR, J Serth, M Schostak *et al.*: DNA methylation biomarkers of prostate cancer: Confirmation of candidates and evidence urine is the most sensitive body fluid for non-invasive detection. *Prostate* (2009).
31. Woodson, K, KJ O'Reilly, JC Hanson, D Nelson, EL Walk,JA Tangrea: The usefulness of the detection of GSTP1 methylation in urine as a biomarker in the diagnosis of prostate cancer. *J Urol* 179, 508-11; discussion 511-2 (2008).
32. Yu, J, T Zhu, Z Wang *et al.*: A novel set of DNA methylation markers in urine sediments for sensitive/specific detection of bladder cancer. *Clin Cancer Res* 13, 7296-304 (2007).
33. Friedrich, MG, DJ Weisenberger, JC Cheng *et al.*: Detection of methylated apoptosis-associated genes in urine sediments of bladder cancer patients. *Clin Cancer Res* 10, 7457-65 (2004).
34. Fiegl, H, C Gattringer, A Widschwendter *et al.*: Methylated DNA collected by tampons--a new tool to detect endometrial cancer. *Cancer Epidemiol Biomarkers Prev* 13, 882-8 (2004).

35. Kahn, SL, BM Ronnett, PE Gravitt,KS Gustafson: Quantitative methylation-specific PCR for the detection of aberrant DNA methylation in liquid-based Pap tests. *Cancer* 114, 57-64 (2008).

36. Olesen, LH, A Aggerholm, BL Andersen *et al.*: Molecular typing of adult acute myeloid leukaemia: significance of translocations, tandem duplications, methylation, and selective gene expression profiling. *Br J Haematol* 131, 457-67 (2005).

37. Strathdee, G, TL Holyoake, A Sim *et al.*: Inactivation of HOXA genes by hypermethylation in myeloid and lymphoid malignancy is frequent and associated with poor prognosis. *Clin Cancer Res* 13, 5048-55 (2007).

38. Widschwendter, M, S Apostolidou, E Raum *et al.*: Epigenotyping in peripheral blood cell DNA and breast cancer risk: a proof of principle study. *PLoS ONE* 3, e2656 (2008).

39. Korshunova, Y, RK Maloney, N Lakey *et al.*: Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. *Genome Res* 18, 19-29 (2008).

40. Devos, T, R Tetzner, F Model *et al.*: Circulating Methylated SEPT9 DNA in Plasma Is a Biomarker for Colorectal Cancer. *Clin Chem* (2009).

41. Grutzmann, R, B Molnar, C Pilarsky *et al.*: Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. *PLoS ONE* 3, e3759 (2008).

42. Lofton-Day, C, F Model, T Devos *et al.*: DNA methylation biomarkers for blood-based colorectal cancer screening. *Clin Chem* 54, 414-23 (2008).

43. Issa, JP: CpG island methylator phenotype in cancer. *Nat Rev Cancer* 4, 988-93 (2004).

44. Hegi, ME, AC Diserens, T Gorlia *et al.*: MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med* 352, 997-1003 (2005).

45. Iafrate, AJ,DN Louis: "MGMT for pt mgmt": Is Methylguanine-DNA Methyltransferase Testing Ready for Patient Management? *J Mol Diagn* 10, 308-10 (2008).

46. Mikeska, T, C Bock, O El-Maarri *et al.*: Optimization of Quantitative MGMT Promoter Methylation Analysis Using Pyrosequencing and Combined Bisulfite Restriction Analysis. *J Mol Diagn* 9, 368-81 (2007).

47. Bernstein, BE, A Meissner,ES Lander: The mammalian epigenome. *Cell* 128, 669-81 (2007).

48. Jeddeloh, JA, JM Greally,OJ Rando: Reduced-representation methylation mapping. *Genome Biol* 9, 231 (2008).

49. Meissner, A, TS Mikkelsen, H Gu *et al.*: Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454, 766-70 (2008).

50. Down, TA, VK Rakyan, DJ Turner *et al.*: A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol* 26, 779-85 (2008).

51. Irizarry, RA, C Ladd-Acosta, B Carvalho *et al.*: Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res* 18, 780-90 (2008).

52. Oda, M, JL Glass, RF Thompson *et al.*: High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res* (2009).

53. Ball, MP, JB Li, Y Gao *et al.*: Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* 27, 361-8 (2009).

54. Deng, J, R Shoemaker, B Xie *et al.*: Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol* 27, 353-60 (2009).

55. Eckhardt, F, J Lewin, R Cortese *et al.*: DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38, 1378-85 (2006).

56. Meng, H, EL Murrelle,G Li: Identification of a small optimal subset of CpG sites as bio-markers from high-throughput DNA methylation profiles. *BMC Bioinformatics* 9, 457 (2008).

57. Tost, J,IG Gut: DNA methylation analysis by pyrosequencing. *Nat Protoc* 2, 2265-75 (2007).

58. Gonzalgo, ML,G Liang: Methylation-sensitive single-nucleotide primer extension (Ms-SNuPE) for quantitative measurement of DNA methylation. *Nat Protoc* 2, 1931-6 (2007).

59. Brena, RM, H Auer, K Kornacker,C Plass: Quantification of DNA methylation in electrofluidics chips (Bio-COBRA). *Nat Protoc* 1, 52-8 (2006).

60. Ehrich, M, MR Nelson, P Stanssens *et al.*: Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proc Natl Acad Sci U S A* 102, 15785-90 (2005).
61. Eads, CA, KD Danenberg, K Kawakami *et al.*: MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res* 28, E32 (2000).
62. Herman, JG, JR Graff, S Myohanen, BD Nelkin,SB Baylin: Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A* 93, 9821-6 (1996).
63. Zhang, Y, C Rohde, S Tierling *et al.*: DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS Genet* 5, e1000438 (2009).
64. Bock, C, J Walter, M Paulsen,T Lengauer: Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Res* 36, e55 (2008).
65. Bock, C,T Lengauer: Computational epigenetics. *Bioinformatics* 24, 1-10 (2008).
66. Reich, M, T Liefeld, J Gould, J Lerner, P Tamayo,JP Mesirov: GenePattern 2.0. *Nat Genet* 38, 500-1 (2006).
67. Gentleman, RC, VJ Carey, DM Bates *et al.*: Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5, R80 (2004).
68. Langmead, B, C Trapnell, M Pop,SL Salzberg: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25 (2009).
69. Li, H, J Ruan,R Durbin: Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18, 1851-8 (2008).
70. Pelizzola, M, Y Koga, AE Urban *et al.*: MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome Res* 18, 1652-9 (2008).
71. Bock, C, K Halachev, J Büch,T Lengauer: EpiGRAPH: User-friendly software for statistical analysis and prediction of (epi-) genomic data. *Genome Biol* 10, R14 (2009).
72. Ji, H, H Jiang, W Ma, DS Johnson, RM Myers,WH Wong: An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 26, 1293-300 (2008).
73. Blankenberg, D, J Taylor, I Schenck *et al.*: A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res* 17, 960-4 (2007).
74. Du, P, WA Kibbe,SM Lin: lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24, 1547-8 (2008).
75. Dunning, MJ, ML Smith, ME Ritchie,S Tavare: beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics* 23, 2183-4 (2007).
76. Frank, E, M Hall, L Trigg, G Holmes,IH Witten: Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479-81 (2004).
77. Bock, C, J Walter, M Paulsen,T Lengauer: CpG island mapping by epigenome prediction. *PLoS Comput Biol* 3, e110 (2007).
78. Bock, C, M Paulsen, S Tierling, T Mikeska, T Lengauer,J Walter: CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet* 2, e26 (2006).
79. Keshet, I, Y Schlesinger, S Farkash *et al.*: Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat Genet* 38, 149-53 (2006).
80. Goh, L, SK Murphy, S Muhkerjee,TS Furey: Genomic sweeping for hypermethylated genes. *Bioinformatics* 23, 281-8 (2007).
81. McCabe, MT, EK Lee,PM Vertino: A multifactorial signature of DNA sequence and polycomb binding predicts aberrant CpG island methylation. *Cancer Res* 69, 282-91 (2009).
82. Rhodes, DR, S Kalyana-Sundaram, V Mahavisno *et al.*: Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9, 166-80 (2007).
83. Lauss, M, I Visne, A Weinhaeusel, K Vierlinger, C Noehammer,A Kriegner: MethCancerDB--aberrant DNA methylation in human cancer. *Br J Cancer* 98, 816-7 (2008).
84. Suderman, M,M Hallett: Tools for visually exploring biological networks. *Bioinformatics* 23, 2651-9 (2007).

85. Tusnady, GE, I Simon, A Varadi,T Aranyi: BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic Acids Res* 33, e9 (2005).

86. Bock, C, S Reither, T Mikeska, M Paulsen, J Walter,T Lengauer: BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics* 21, 4067-8 (2005).

87. Kumaki, Y, M Oda,M Okano: QUMA: quantification tool for methylation analysis. *Nucleic Acids Res* 36, W170-5 (2008).

88. Thompson, RF, M Suzuki, KW Lau,JM Greally: A pipeline for the quantitative analysis of CG dinucleotide methylation using mass spectrometry. *Bioinformatics* (2009).

89. Tarca, AL, VJ Carey, XW Chen, R Romero,S Draghici: Machine learning and its applications to biology. *PLoS Comput Biol* 3, e116 (2007).

90. Witten, IH,E Frank: Data mining : practical machine learning tools and techniques with Java implementations. 2000, San Francisco, Calif.: Morgan Kaufmann. xxv, 371 p.

91. Jurisica, I, D Wigle,B Wong: Cancer informatics in the post genomic era: toward information-based medicine. 2007, New York ; London: Springer. xix, 180 p.

92. Bibikova, M,JB Fan: GoldenGate assay for DNA methylation profiling. *Methods Mol Biol* 507, 149-63 (2009).

93. Bibikova, M, Z Lin, L Zhou *et al.*: High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* 16, 383-93 (2006).

94. Ley, TJ, ER Mardis, L Ding *et al.*: DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456, 66-72 (2008).

95. Wood, LD, DW Parsons, S Jones *et al.*: The genomic landscapes of human breast and colorectal cancers. *Science* 318, 1108-13 (2007).

96. Altshuler, D, MJ Daly,ES Lander: Genetic mapping in human disease. *Science* 322, 881-8 (2008).

97. Ludwig, JA,JN Weinstein: Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer* 5, 845-56 (2005).

98. Cancer Genome Atlas Research Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061-8 (2008).

99. Herman, JG,SB Baylin: Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* 349, 2042-54 (2003).

100. Jones, PA,SB Baylin: The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 3, 415-28 (2002).

101. Chan, TA, S Glockner, JM Yi *et al.*: Convergence of mutation and epigenetic alterations identifies common genes in cancer that predict for poor prognosis. *PLoS Med* 5, e114 (2008).

102. Hong, C, KS Moorefield, P Jun *et al.*: Epigenome scans and cancer genome sequencing converge on WNK2, a kinase-independent suppressor of cell growth. *Proc Natl Acad Sci U S A* 104, 10974-9 (2007).

103. Schuebel, KE, W Chen, L Cope *et al.*: Comparing the DNA hypermethylome with gene mutations in human colorectal cancer. *PLoS Genet* 3, 1709-23 (2007).

104. Pepe, MS, R Etzioni, Z Feng *et al.*: Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst* 93, 1054-61 (2001).

105. Johannes, F, V Colot,RC Jansen: Epigenome dynamics: a quantitative genetics perspective. *Nat Rev Genet* 9, 883-90 (2008).

106. Butcher, LM,S Beck: Future impact of integrated high-throughput methylome analyses on human health and disease. *J Genet Genomics* 35, 391-401 (2008).

107. Foley, DL, JM Craig, R Morley *et al.*: Prospects for epigenetic epidemiology. *Am J Epidemiol* 169, 389-400 (2009).

108. McCarthy, MI, GR Abecasis, LR Cardon *et al.*: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9, 356-69 (2008).

109. Becker, KG, DA Hosack, G Dennis, Jr. *et al.*: PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* 4, 61 (2003).

Reference annotations

[9] – The authors sketch a conceptual framework explaining the interplay of genetic and epigenetic factors in complex diseases.
[44] – A multi-center clinical trial confirms the utility of *MGMT* promoter methylation for predicting chemotherapy resistance in brain tumors.
[105] – The authors discuss concepts for the integration of epigenetic analysis with association studies and linkage analysis.
[40-42] – Several validation studies confirm *septin 9* hypermethylation as a blood-based epigenetic biomarker for diagnosing colon cancer.
[26] – The authors identify a genetically and epigenetically distinct subgroup of colon cancer patients and devise a discriminatory biomarker.

All six references are to be tagged as "of interest (*)".

Figures

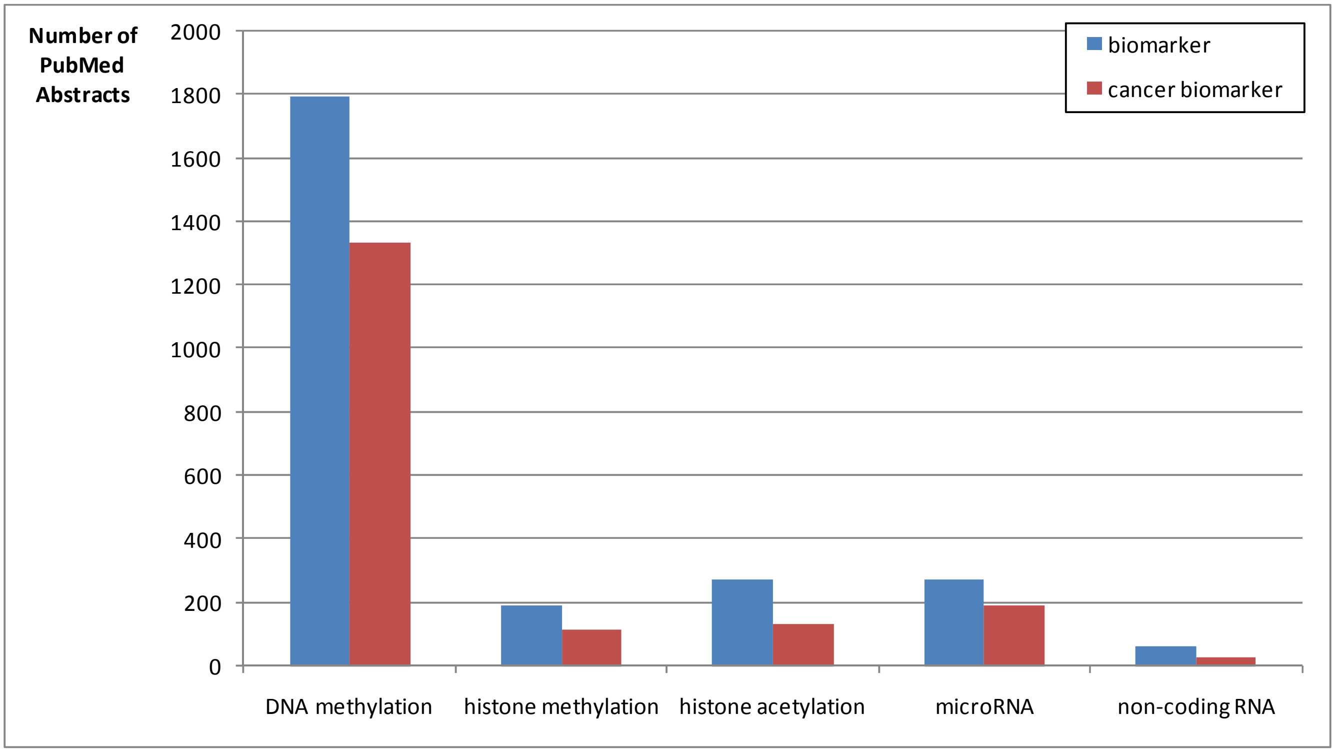
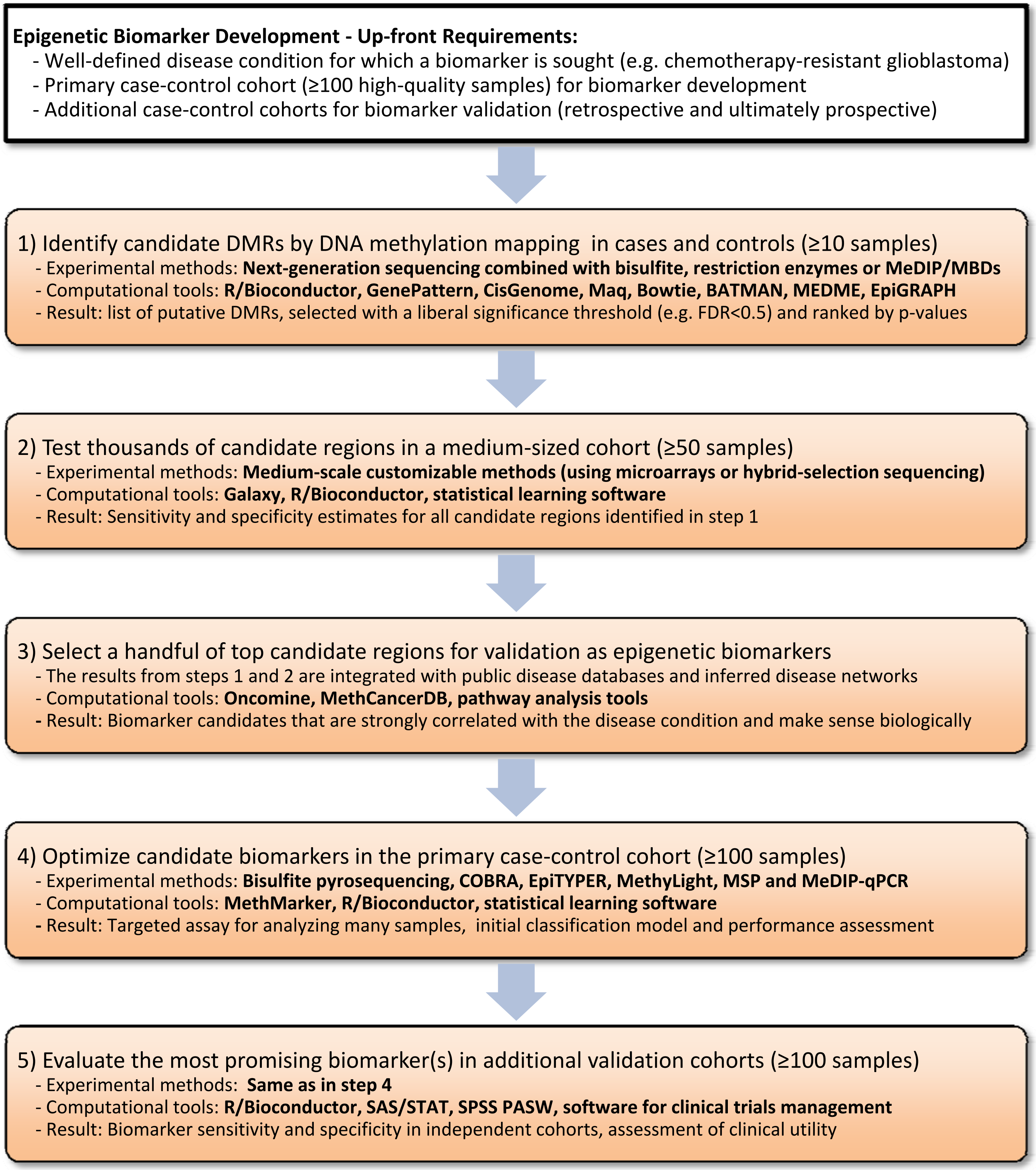


Figure 1. DNA methylation is widely used as an epigenetic biomarker

The diagram visualizes the number of PubMed abstracts in which the terms "biomarker" (blue) or "cancer biomarker" (red) co-occur with one of several terms representing epigenetic phenomena (on the x-axis). Citation data were retrieved with the help of PubMatrix [109] in June 2009.



Epigenetic Biomarker Development - Up-front Requirements:

- Well-defined disease condition for which a biomarker is sought (e.g. chemotherapy-resistant glioblastoma)
- Primary case-control cohort (≥100 high-quality samples) for biomarker development
- Additional case-control cohorts for biomarker validation (retrospective and ultimately prospective)

1) Identify candidate DMRs by DNA methylation mapping in cases and controls (≥10 samples)

- Experimental methods: **Next-generation sequencing combined with bisulfite, restriction enzymes or MeDIP/MBDs**
- Computational tools: **R/Bioconductor, GenePattern, CisGenome, Maq, Bowtie, BATMAN, MEDME, EpiGRAPH**
- Result: list of putative DMRs, selected with a liberal significance threshold (e.g. FDR<0.5) and ranked by p-values

2) Test thousands of candidate regions in a medium-sized cohort (≥50 samples)

- Experimental methods: **Medium-scale customizable methods (using microarrays or hybrid-selection sequencing)**
- Computational tools: **Galaxy, R/Bioconductor, statistical learning software**
- Result: Sensitivity and specificity estimates for all candidate regions identified in step 1

3) Select a handful of top candidate regions for validation as epigenetic biomarkers

- The results from steps 1 and 2 are integrated with public disease databases and inferred disease networks
- Computational tools: **Oncomine, MethCancerDB, pathway analysis tools**
- Result: Biomarker candidates that are strongly correlated with the disease condition and make sense biologically

4) Optimize candidate biomarkers in the primary case-control cohort (≥100 samples)

- Experimental methods: **Bisulfite pyrosequencing, COBRA, EpiTYPER, MethyLight, MSP and MeDIP-qPCR**
- Computational tools: **MethMarker, R/Bioconductor, statistical learning software**
- Result: Targeted assay for analyzing many samples, initial classification model and performance assessment

5) Evaluate the most promising biomarker(s) in additional validation cohorts (≥100 samples)

- Experimental methods: **Same as in step 4**
- Computational tools: **R/Bioconductor, SAS/STAT, SPSS PASW, software for clinical trials management**
- Result: Biomarker sensitivity and specificity in independent cohorts, assessment of clinical utility

Figure 2. A systematic approach to epigenetic biomarker discovery

This diagram outlines a procedure designed to guide and accelerate the development of DNA methylation biomarkers. Bioinformatic tools are highlighted that support and partially automate key tasks in each phase. All listed software packages are freely available for academic use. Abbreviations: DMRs – Differentially Methylated Regions; MeDIP – Methylated DNA Immunoprecipitation; MBDs – Methyl-CpG Binding Domains; FDR – False Discovery Rate; COBRA – Combined Bisulfite Restriction Analysis; MSP – Methylation Specific PCR.