BLUEPRINT to decode the epigenetic signature written in blood

To the Editor:

Last October, scientists gathered in Amsterdam to celebrate the start of BLUEPRINT (http://www.blueprint-epigenome.eu/), an EU-funded consortium that will generate epigenomic maps of at least 100 different blood cell types. With this initiative, Europe has pledged a substantial contribution to the ultimate goal of the International Human Epigenome Consortium (IHEC) to map 1,000 human epigenomes. Here, we provide a brief background to the scientific questions that prompted the formation of BLUEPRINT, summarize the overall goals of BLUEPRINT and detail the specific areas in which the consortium will focus its initial efforts and resources.

In mammals, nucleated cells share the same genome but have different epigenomes depending on the cell type and many other factors, resulting in an astounding diversity in phenotypic plasticity with respect to morphology and function. This diversity is defined by cell-specific patterns of gene expression, which are controlled through regulatory sites in the genome to which transcription factors bind. In eukaryotes, access to these sites is orchestrated via chromatin, the complex of DNA, RNA and proteins that constitutes the functional platform of the genome. In contrast with DNA, chromatin is not static but highly dynamic, particularly through modifications of histones at nucleosomes and cytosines at the DNA level that together define the epigenome, the epigenetic state of the cell. Advances in new genomics technologies, particularly next-generation sequencing, allow the epigenome to be studied in a holistic fashion, leading to a better understanding of chromatin function and functional annotation of the genome. Yet little is known about how epigenetic characteristics vary between different cell types, in health and disease or among individuals. This lack of a quantitative framework for the dynamics of the epigenome and its determinants is a major hurdle for the translation of epigenetic observations into regulatory models, the identification of associations between epigenotypes and diseases, and the subsequent development of new classes of compounds for disease prevention and treatment. The task, however, is daunting as each of the several hundred cell types in the human body is expected to show specific epigenomic features that are further expected to respond to environmental inputs in time and space. The research community has realized these limitations and the need for concerted action.

The IHEC was founded to coordinate large-scale international efforts toward the goal of a comprehensive human epigenome reference atlas (http://www.ihec-epigenomes.org/). The IHEC will coordinate epigenomic mapping and characterization worldwide to avoid redundant research efforts, implement high data quality standards, coordinate data storage, management and analysis, and provide free access to the epigenomes produced. The maps generated under the umbrella of the IHEC contain detailed information on DNA methylation, histone modification, nucleosome occupancy, and corresponding coding and noncoding RNA expression in different normal and diseased cell types. This will allow integration of different layers of epigenetic information for a wide variety of distinct cell types and thus provide a resource for both basic and applied research.

BLUEPRINT aims to bridge the gap in our current knowledge between individual components of the epigenome and their functional dynamics through state-of-the-art analysis in a defined set of primarily human hematopoietic cells from healthy and diseased individuals. Mammalian blood formation or hematopoiesis is one of the best-studied systems of stem cell biology. Blood formation can be viewed as a hierarchical process, and classically, differentiation is defined to occur along the myeloid and lymphoid lineages. The identity of cellular intermediates and the geometry of branch points are still under intense investigation and therefore provide a paradigm for delineation of fundamental principles of cell fate determination and regulation of proliferation and lifespan, which differ considerably between different types of blood cells.

BLUEPRINT will generate reference epigenomes of at least 50 specific blood cell types and their malignant counterparts and aim to provide high-quality reference epigenomes of primary cells from >60 individuals with detailed genetic and, where appropriate, medical records. To account for and quantify the impact of DNA sequence variation on epigenome differences, BLUEPRINT will work whenever possible on samples of known genetic variation, including samples from the Cambridge BioResource (Cambridge, UK), the International Cancer Genome Consortium and the British Diabetic Twin Study for disease-discordant monozygotic twin samples. The Wellcome Trust Sanger Institute (Hinxton, UK) will also provide full genomic sequencing for up to 100 samples. BLUEPRINT will harness existing proven technologies to generate reference epigenomes, including RNA-Seq for transcriptome analysis, bisulfite sequencing for methylome analysis, DNaseI-Seq for analysis of hypersensitive sites and ChIP-Seq for analysis of at least six histone marks. Moreover, BLUEPRINT aims to develop new technologies to enhance high-throughput epigenome mapping, particularly when using few cells.

BLUEPRINT is initially focusing on four main areas. One main goal of the project is to comprehensively analyze diverse epigenomic maps and make them available as an integrated BLUEPRINT-IHEC resource to the scientific community. Integration is envisioned for related projects within species (e.g., the 1000 Genomes Project) and between species (e.g., modENCODE) to better understand functional aspects (e.g., shared pathways) and the evolution of cell lineage development. Analysis of the BLUEPRINT data is expected to catalyze a better understanding of the relationship between epigenetic and genomic information and will form the basis for generation of new methods (e.g., epigenetic imputation) for prediction of epigenetic states from epigenomic profiles. Such prediction methods will facilitate a move toward a more quantitative knowledge and modeling of epigenetic mechanisms. As a result, such models could in the future assist in 'reverse engineering' of regulatory networks to repair or restore epigenetic codes that have been perturbed by disease.

A second goal of BLUEPRINT is to systematically link epigenetic variation with phenotypic plasticity in health and disease. This will be attempted in three ways. First, genetic and epigenetic varation in two blood cell types from 100 healthy individuals will be analyzed. These measurements will be combined with whole-genome and transcriptome sequencing to dissect the interplay between common DNA sequence

variation and the epigenome. This will allow estimation, assessed by changes in transcription, of the degree to which epigenetic variation is driven by genetic variation and how this variation affects function. Second, epigenetic profiles will be generated from comparable cell types in three different mouse strains for which high-quality sequence is now available. Determining genotype-epigenotype variation in mouse will allow detailed comparison with human data sets, contribute to generation of experimentally testable hypotheses in a tractable model organism and provide a framework for future comparative epigenomic analyses of other purified mouse cell populations. Third, the possible role of epigenetic variation (that is, DNA methylation) in the etiology of common diseases will be determined by the first comprehensive epigenome-wide association study of any human disease. This is designed to discriminate between epigenetic variation that is a cause or a consequence of disease, a currently unresolved issue in epigenome-wide association studies. Type 1 diabetes has been chosen as an exemplar because certain blood cells (e.g., $CD14^+$ monocytes) are relevant to type 1 diabetes pathogenesis and because BLUEPRINT partners have already conducted a successful pilot study demonstrating the involvement of methylation-variable positions, the epigenetic equivalent of single-nucleotide polymorphisms, in type 1 diabetes. Of the >100 methylation-variable positions identified, many have been associated with immune genes and have been found present before disease diagnosis, as well as in patients positive for diabetes-associated autoantibodies but disease-free after 12 years[1]. These findings suggest methylation-variable positions to be involved in type 1 diabetes etiology as they cannot be explained by genetic heterogeneity or twinning, metabolic dysfunction, insulin or other pharmacological treatment.

Another goal of BLUEPRINT is to foster the clinical relevance of epigenetic analysis by including a major effort in the biomarker area. The focus will be identifying biomarkers for more accurate prognosis and personalized therapy of childhood acute lymphoblastic leukemia and for determining the efficacy of epigenetic drug treatment in acute myeloid leukemia and myelodysplastic syndrome. Biomarker development will focus specifically on DNA methylation to maximize compatibility with clinical diagnostics and capitalize on the clinical sequencing infrastructure becoming available in European countries[2].

BLUEPRINT's last goal is to identify new compounds that interact with epigenetic regulators. Epigenetic modifications are reversible and thus have the potential to be modified by small molecule drugs. The first epigenetic targeting drugs (DNA methyltransferase and histone deacetylase (HDAC) inhibitors) have efficacy in some types of cancer but are currently limited by lack of specificity and high toxicity. The full potential of epigenetic-based therapies has not yet been realized owing to limitations of current knowledge on specific targets and imperfect strategies of current drug discovery approaches based on recombinant, single target assays. Although deregulated expression of a few epitargets has been found in cancer, functional results are required to determine which targets qualify as candidates warranting follow-up studies. BLUEPRINT will use RNA interference screens to validate epigenetic targets in *ex vivo* and *in vivo* settings. Particularly relevant is the concept of direct *in vivo* validation of candidate targets, a relatively new approach, on a selected set of candidate drug targets. *In vivo* RNA interference–based screens in murine leukemia models and patient-derived human samples will be carried out to achieve the most relevant level of preclinical validation possible. Cell assays and screens will be complemented by *in vivo* screens, used for further validation and, most importantly, allow further characterization of biological phenotypes and assays to be developed for drug discovery. Thus, the conventional process of sequential validation (from *in vitro* to *in vivo* studies) will be improved to provide a more innovative strategy.

BLUEPRINT will also pioneer approaches to drug discovery based on isolating epigenetic complexes directly from human cells or tissues, thus preserving their native multicomponent structure. This is an essential requirement for focused target identification and compound optimization to produce clinically relevant therapeutics. For instance, by using beads modified with HDAC inhibitors to capture interacting proteins from cell lysates and quantitative mass spectrometry to determine drug affinities for those complexes, Cellzome (Heidelberg, Germany) has already identified novel histone deacetylase–protein complexes specific for defined cell states and several new targets for HDAC inhibitors[3]. By screening existing, non-HDAC inhibitors for binding to HDAC complexes, potential new uses for these drugs as epigenetic modulators can be uncovered, allowing either direct application as epigenetic therapy or as a new starting point for lead optimization on HDAC targets. Additionally, the Cellzome platform has revealed unexpected selectivity of known HDAC clinical-stage inhibitors and lack of activity against specific HDAC protein complexes. We think many of these basic principles can be extended to other epigenetic target classes.

In summary, the commencement of BLUEPRINT signals a new era for European biomedical research that we hope will harness the power of epigenomic dynamics in health and disease. We expect the results will contribute to our knowledge of genome biology and facilitate the definition of new avenues for pharmaceutical intervention, prediction and diagnosis of disease.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology.

*The BLUEPRINT Consortium**

***David Adams[1], Lucia Altucci[2], Stylianos E Antonarakis[3], Juan Ballesteros[4], Stephan Beck[5], Adrian Bird[6], Christoph Bock[7], Bernhard Boehm[8], Elias Campo[9], Andrea Caricasole[10], Fredrik Dahl[11], Emmanouil T Dermitzakis[3], Tariq Enver[5], Manel Esteller[12], Xavier Estivill[13], Anne Ferguson-Smith[14], Jude Fitzgibbon[15], Paul Flicek[16], Claudia Giehl[17], Thomas Graf[13], Frank Grosveld[18], Roderic Guigo[13], Ivo Gut[19], Kristian Helin[20], Jonas Jarvius[21], Ralf Küppers[22], Hans Lehrach[23], Thomas Lengauer[7], Åke Lernmark[24], David Leslie[15], Markus Loeffler[25], Elizabeth Macintyre[26], Antonello Mai[27], Joost HA Martens[28], Saverio Minucci[29], Willem H Ouwehand[14], Pier Giuseppe Pelicci[29], Hélène Pendeville[30], Bo Porse[20], Vardhman Rakyan[15], Wolf Reik[31], Martin Schrappe[32], Dirk Schübeler[33], Martin Seifert[32], Reiner Siebert[34], David Simmons[35], Nicole Soranzo[1], Salvatore Spicuglia[36], Michael Stratton[1], Hendrik G Stunnenberg[28], Amos Tanay[37], David Torrents[38], Alfonso Valencia[39], Edo Vellenga[40], Martin Vingron[23], Jörn Walter[41] & Spike Willcocks[42]*

[1]Wellcome Trust Sanger Institute, Hinxton, UK. [2]Second University of Naples, Naples, Italy. [3]Universite De Geneve, Geneva, Switzerland. [4]Vivia Biotech, Madrid, Spain. [5]University College London, London, UK. [6]University of Edinburgh, Edinburgh, UK. [7]Max Planck Institute for Informatics, Saarbrücken, Germany. [8]University of Ulm, Ulm, Germany. [9]Institut D'investigacions Biomediques August Pi I Sunyer, Barcelona, Spain. [10]Siena Biotech Spa, Siena, Italy. [11]Halo Genomics AB, Uppsala, Sweden. [12]Institut D'investigacio Biomedica De

Bellvitge, Barcelona, Spain. [13]Centre for Genomic Regulation, Barcelona, Spain. [14]University of Cambridge, Cambridge, UK. [15]Queen Mary University of London, London, UK. [16]European Bioinformatics Institute, Hinxton, UK. [17]European Research and Project Office GmbH, Saarbrücken, Germany. [18]Erasmus University Medical Centre Rotterdam, Rotterdam, The Netherlands. [19]Centro Nacional de Analisis Genomico, Barcelona, Spain. [20]University of Copenhagen, Copenhagen, Denmark. [21]Sigolis AB, Uppsala, Sweden. [22]Universitaetsklinikum Essen, Essen, Germany. [23]Max Planck Institute for Molecular Genetics, Berlin, Germany. [24]Lund University, Malmo, Sweden. [25]University of Leipzig, Leipzig, Germany. [26]Centre National de la Recherche Scientifique, Paris Descartes University, Paris, France. [27]Sapienza University of Rome, Rome, Italy. [28]Radboud University, Nijmegen, The Netherlands. [29]European Institute of Oncology, Milan, Italy. [30]Diagenode SA, Liege, Belgium. [31]The Babraham Institute, Cambridge, UK. [32]Genomatix Software GmbH, Munich, Germany. [33]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. [34]Christian-Albrechts-Universitaet Zu Kiel, Kiel, Germany. [35]Cellzome AG, Heidelberg, Germany. [36]Institut National de la Sante et de la Recherche Medicale, Marseille, France. [37]Weizmann Institute of Science, Rehovot, Israel. [38]Barcelona Supercomputing Center, Barcelona, Spain. [39]Centro Nacional de Investigaciones Oncologicas, Madrid, Spain. [40]University Medical Centre Groningen, Groningen, The Netherlands. [41]University of Saarland, Saarbruecken, Germany. [42]Oxford Nanopore Technologies Ltd., Oxford, UK.
e-mail: h.stunnenberg@ncmls.ru.nl

1. Rakyan, V.K. *et al. PLoS Genet.* **9**, e1002300 (2011).
2. Callaway, E. *Nature* **467**, 766–767 (2010).
3. Bantscheff, M. *et al. Nat. Biotechnol.* **29**, 255–265 (2011).

Detecting and annotating genetic variations using the HugeSeq pipeline

To the Editor:

Deciphering genome sequences is important for the mapping of genetic diseases and prediction of their risks. Advances in high-throughput DNA sequencing technologies using short read lengths have enabled rapid sequencing of entire human genomes and unlocked the potential for comprehensive identification of their underlying genetic variations. Various computational algorithms for identifying and characterizing variants have been developed; however, most of these computational methods are neither integrated nor interoperable, making it difficult for biologists to extract all the genetic information from billions of sequences generated by these sequencing technologies. Here, we present HugeSeq, an integrated computational pipeline to fully automate the process of variant detection from alignment of these genomic sequences to detection and annotation of all types of genetic variations (single nucleotide polymorphisms (SNPs), short insertions or deletions (indels) and larger structural variations (SVs)).

Compared with other popular platforms for genome data analysis that typically analyze SNPs or a limited set of variants (**Supplementary Table 1**), HugeSeq covers a more complete spectrum of variant types. The complete variant detection and characterization workflow of the HugeSeq pipeline (**Fig. 1**) is a modular framework comprising three phases: first, a mapping phase that prepares and aligns reads; second, a sorting phase that combines and sorts alignments for parallel variant detection; and third, a reduction phase that detects and annotates different variants (SNPs, indels and structural variations). It is based on a MapReduce[1] approach and runs in a parallel computational environment, making it highly efficient and scalable.

HugeSeq uses sequence reads (both single end and paired end) in a FASTA or FASTQ format (optionally compressed in a GZIP format) as input for alignment. Because alignment of a single read is independent of others, HugeSeq divides the reads into smaller subsets so they can be aligned in parallel. It then distributes the reads in the computer cluster and carries out a gapped alignment against the reference genome using the Burrows-Wheeler aligner[2]. The generated sequence alignment map (SAM)[3] is then converted into its binary format, BAM, using SAMtools[3] to ensure efficient storing and access of alignment information. After alignment, HugeSeq collects all the mapped reads and sorts them according to their aligned chromosomal positions with the Picard tool. The sorted reads for each chromosome are assigned to their corresponding chromosomal BAM and indexed for rapid and random access using SAMtools. Because most variant detections are intrachromosomal, the detection process can be carried out on each chromosomal BAM simultaneously. Interchromosomal translocation detection can also be enabled and run in a nonparallel mode, although it slows down the process considerably.

To enhance the quality of the alignments for more accurate variant detection, HugeSeq carries out several processing ('cleanup') procedures before variant calling. First, to minimize experimental artifacts, it removes potential PCR duplicates using the Picard tool. Second, it carries out a local realignment around indels and SNP clusters using the Genome Analysis ToolKit (GATK) realigner[4]. Last, based on the realignment, it recalibrates the base quality of the alignments using the GATK recalibrater[4] so that the quality scores represent the empirical probability of mismatching to the reference genome. With the processed read alignments or any user-specified BAMs, HugeSeq detects variants of different kinds in a parallel fashion.

For SNP and small indel detection, HugeSeq uses two different well-established SNP and indel calling algorithms, the GATK UnifiedGenotyper[4] and SAMtools[3]. When calling indels using GATK, it uses the Dindel[5] model for greater sensitivity. The resulting SNPs and indels are then passed through the GATK variant filtering tool with default parameters similar to those used in the 1000 Genomes Project[6]. Structural variations and copy number variants (CNVs) are often difficult to detect, largely owing to their heterogeneous nature. A variety of different methods can be used to find them but each has distinct biases. To identify as many structural variations as possible, HugeSeq uses four major approaches: first, paired-end mapping using BreakDancer[7]; second, split-read analysis using Pindel[8]; third, read-depth analysis using CNVnator[9] and fourth, junction mapping using BreakSeq[10] (a version we modified to support BAM as input for unmapped reads). Because these structural variation and CNV callers generate variant calls in different formats, HugeSeq standardizes their outputs by converting them into the standard general feature format.

The resulting SNP and indel call sets, which are in a standard variant call format (VCF), are combined and merged using VCFtools[11]. HugeSeq also uses VCFtools to concatenate variants from different BAMs for each algorithm and to merge calls from different algorithms into a single VCF. SNPs called by both GATK and SAMtools are of particularly high confidence. For the structural variation