

RESEARCH ARTICLE

Open Access



Automatic detection of genomic regions with informative epigenetic patterns

Florencio Pazos^{*} , Adrian Garcia-Moreno and Juan C. Oliveros

Abstract

Background: Epigenetic phenomena are crucial for explaining the phenotypic plasticity seen in the cells of different tissues, developmental stages and diseases, all holding the same DNA sequence. As technology is allowing to retrieve epigenetic information in a genome-wide fashion, massive epigenomic datasets are being accumulated in public repositories. New approaches are required to mine those data to extract useful knowledge. We present here an automatic approach for detecting genomic regions with epigenetic variation patterns across samples related to a grouping of these samples, as a way of detecting regions functionally associated to the phenomenon behind the classification.

Results: We show that the regions automatically detected by the method in the whole human genome associated to three different classifications of a set of epigenomes (cancer vs. healthy, brain vs. other organs, and fetal vs. adult tissues) are enriched in genes associated to these processes.

Conclusions: The method is fully automatic and can exhaustively scan the whole human genome at any resolution using large collections of epigenomes as input, although it also produces good results with small datasets. Consequently, it will be valuable for obtaining functional information from the incoming epigenomic information as it continues to accumulate.

Keywords: Epigenomics, Epigenetics, Gene transcription regulation

Background

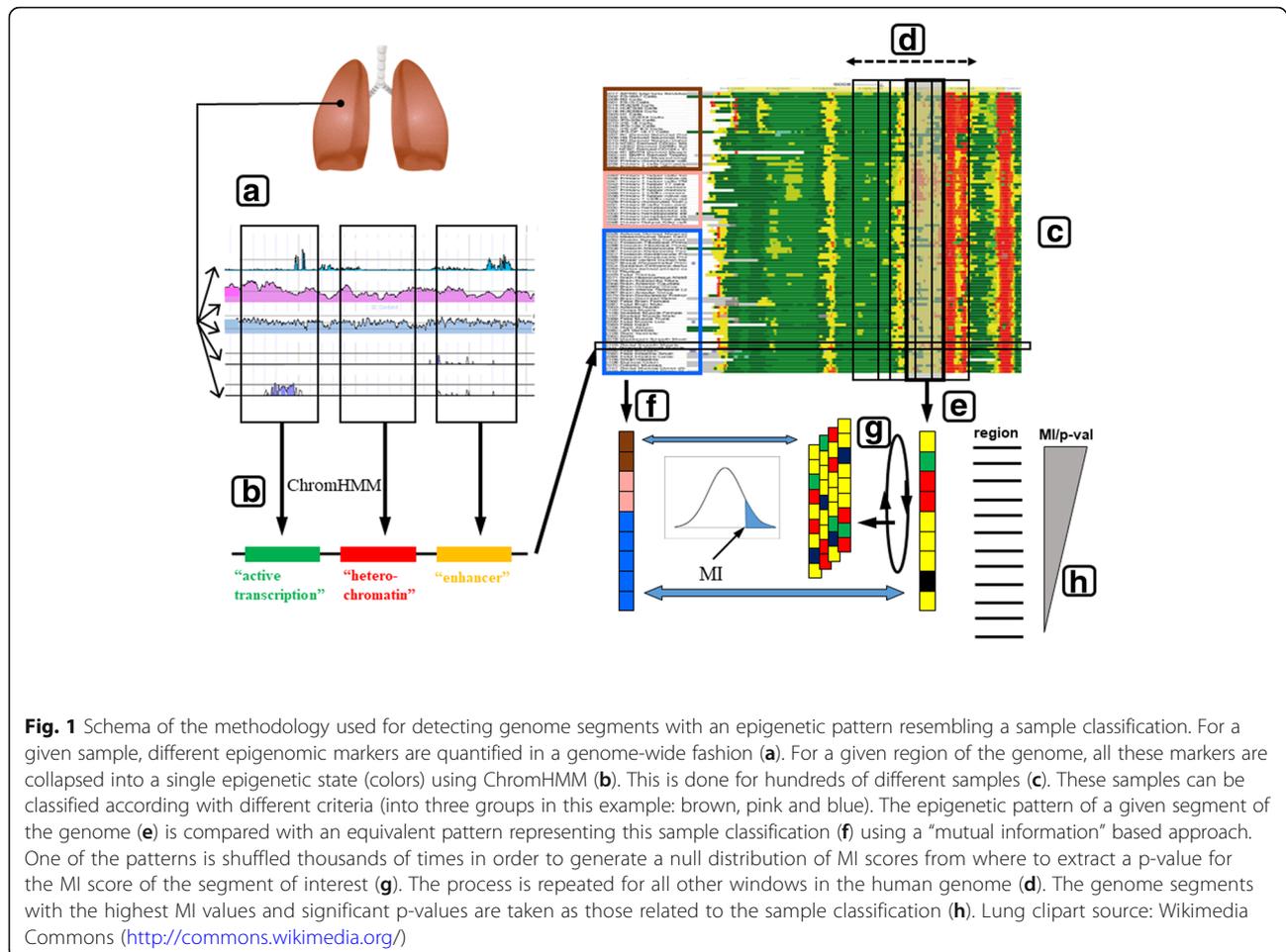
The expression of the information encoded in the genome is determined not only by its DNA sequence but by many other factors that affect the complex process of gene expression [1]. These factors include reversible modifications in the DNA and histones that affect gene expression without altering the nucleotide sequence of the DNA, for example by changing the folding state or accessibility of the chromatin. These are called epigenetic modifications, and “epigenomics” is the high-throughput characterization of these modifications at a genomic scale [2, 3]. Epigenetic control of gene expression has been shown to be crucial in processes such as development and cell differentiation [4, 5], as well as in diseases [6, 7]. In the case of diseases, epigenetic mechanisms are gaining increasing attention since they could explain the influence of environmental factors in a pathological state [8].

Using assays based on deep sequencing, it is possible to detect in a genome-wide fashion different epigenetic modifications (i.e. covalent modifications of DNA and histones), as well as other proxies of the epigenetic state of a genome segment (e.g. mRNA expression, nucleosome occupancy), for a given sample (Fig. 1). Recently, different consortia at national and international levels were assembled with the objective of obtaining this genome-wide epigenomic information for hundreds of samples (“epigenomes”), including different organs, developmental states and cancer lines [3]. Examples of such initiatives are ENCODE [9], BLUEPRINT [10] and Roadmap Epigenomics [11]. This massive epigenomic information is starting to be mined in a similar way as genomic data. For example, it is being used, alone or in combination with other evidences, to reconstruct the 3D structure of the chromatin [12–14], detect functionally related genes [15], interpreting the results of “genome-wide association studies” (GWAS) [16] and non-coding variants in general [17]. This genome-wide epigenetic information is also being used for epigenomic-GWAS

* Correspondence: pazos@cnb.csic.es

National Center for Biotechnology (CNB-CSIC), c/ Darwin, 3, 28049 Madrid, Spain





(EGWAS) [18], to establish links between epigenetic variation in certain loci and diseases. Recently, it was also used to look for regions whose quantitative epigenetic marks change between two sets of samples, an strategy implemented, for example, in the ChromSwitch package [19].

From a practical point of view, all those different indicators of epigenetic state for a given genome segment can be collapsed into a single “epigenetic state” easier to be interpreted (e.g. “active transcription”, “enhancer”, ...) (Fig. 1). This is done by training machine-learning systems with segments of known epigenetic state and their corresponding profiles of epigenetic proxies. Once trained, these methods are used to infer the state for a given region based on its raw epigenetic profiles. Examples of these methodologies are ChromHMM [20] and Segway [21].

These datasets of chromatin states are conceptually similar to protein or DNA multiple sequence alignments (MSAs): the epigenomes would be the homologous proteins and the genome positions the residues (Fig. 1). Indeed, methodologies originally developed for protein and DNA alignments are being adapted to these datasets: e.g.

Epilogos (<https://epilogos.altiusinstitute.org/>) for generating state-logos (equivalent to sequence logos), and ChromDet [22] for the unsupervised classification of epigenomes and the detection of chromatin regions associated to that classification (CDRs: “chromatin determining regions”). These are equivalent to the “specificity determining residues” (SDPs) in protein MSAs: residues differentially conserved in the subfamilies of the MSA and hence responsible for their functional specificity [23–25].

Nevertheless, such a rich source of information is still scarcely explored, compared with the thousands of approaches available for studying, for example, DNA sequences (“genomics”). Apart from providing interesting information on particular genes and cellular processes, the large scale analysis of these epigenomic datasets might provide systemic information. Moreover, while all studies report a large global epigenetic variation contributing to cell differentiation, locating where the epigenetic changes responsible for it exactly occur is still a challenge.

In this work we propose a methodology for mining the growing epigenomic data in the search for functional

genomic regions. This approach automatically detects genomic regions whose epigenetic profile across a set of samples is related to a user-provide classification of these samples. For the regions detected in this way that contain coding genes, we show that these are involved in the biological processes associated to the sample classification. We also show that the method is robust to the lack of input epigenomic information and can work with a reduced number of epigenomes, although it is fast enough for handling the expected increasing number of epigenomes in the future.

Methods

In order to find genomic regions whose inter-sample epigenomic variation correlates with a given classification of these, we used the epigenomic datasets compiled by the Roamap Epigenomics Consortium [11]. We downloaded the 127 “consolidated epigenomes”, for which the chromatin states are given in a 15-state vocabulary, from https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html

These 127 epigenomes include different tissues, primary cells, cancer cell lines, different developmental stages, etc. For each epigenome, detailed information on different proxies of the epigenetic state at 1 bp resolution is provided, such as histone acetylation, DNA methylation, DNase accessibility or RNA-seq expression (“a” in Fig. 1). For each 200 bp genomic region, the ChromHMM method [20] was used to collapse all these experimental epigenetic proxies into a single “epigenetic state” with 15 possible values (“b” in Fig. 1). From the URL above, we downloaded the BED files with the mnemonics indicating the chromatin states along the 127 epigenomes at that 200 bp resolution (“c” in Fig. 1). The original 15-state vocabulary can be simplified by grouping states representing similar epigenetic behavior. Apart from the original 15 states, for this work we also tried two reductions (to 5 and 2 states –i.e. “active” vs. “inactive”-) previously used by [22].

In order to scan the human genome for segments with a given pattern of epigenetic changes, we use a sliding window of a given length (200 bp or larger) and move it at different steps along the whole genome (“d” in Fig. 1). The epigenetic state for that window in a given epigenome is taken as that with the highest frequency among all 200 bp segments within the window. A minimum frequency of 80% is required. Otherwise, the “undefined” state is assigned to that window. So, for a given window, we end up in a vector (epigenetic profile) where each component represents an epigenome and its value is the epigenetic state at that particular genomic region (with 15, 5 or 2 possible states, depending on the vocabulary used, plus “undefined”) (“e” in Fig. 1). An equivalent vector is constructed representing the classification of the samples in a given number of groups (“f” in Fig. 1). The

similarity between these two vectors is quantified by their Mutual Information. The mutual information between the group classification vector (G) and the epigenetic profile of the i_{th} segment (S_i) is calculated as:

$$MI(G : S_i) = \sum_{k=1}^n \sum_{l=1}^m P(G_k, S_{il}) \cdot \log_2 \frac{P(G_k, S_{il})}{P(G_k) \cdot P(S_{il})}$$

Where the sums run for all possible groups in the sample classification (n) and all possible epigenetic states (m). $P(G_k)$ is the frequency of group k , $P(S_{il})$ is the frequency of the epigenetic state l in window i , and $P(G_k, S_{il})$ the frequency of group k matching state l . State “undefined” is ignored for this calculation, and the terms of the sums where either $P(G_k)$ or $P(S_{il})$ are 0 are skipped. Consequently, this parameter has a high value for a window when its epigenetic profile correlates with the sample classification. That is, all samples of the same group tend to present the same epigenetic state, and the other way around. In that case there is a relationship between the epigenetic states and the group classification.

To assess the significance of an observed MI value we shuffle the components of epigenetic profile vector of the corresponding window a number of times (10,000 in this work) and calculate the distribution of MI values of the sample vector against these random vectors. The composition of the window vector is preserved in this randomization, and only the positions of the labels are changed (“g” in Fig. 1). From this background distribution, a p -value for a given observed MI score is calculated as the fraction of randomized vectors resulting in higher or equal MI’s. These p -values are additionally adjusted for multiple testing using Benjamini-Hochberg’s FDR correction. So, at the end, for a given group classification, we end up with a list of all genome segments (windows) and their associated MI score and p -values (“h” in Fig. 1).

To summarize the epigenetic change pattern of a given window across a large number of samples classified in groups, we just take the majority state for each group.

To show the potential of this methodology to locate genomic regions associated to different phenomena, we applied it to three different classifications of subsets the 127 epigenomes:

- Brain vs. other tissues. We filtered out samples from cultured cells, cancer-derived and those annotated as “other”. Our final list contains 75 epigenomes from which 10 belong to the “brain” group (including different brain anatomical parts, male/female, fetal/adult, etc.)
- Cancer vs. normal. We took the five cancer epigenomes available and all samples of the corresponding healthy tissues. So, we do not include a sample of a given

tissue if there is no cancer version for it. The final list contains 21 epigenomes (5 cancer).

- Fetal vs. other. We filtered out samples from cultured cells and cancer-derived. The final list contains 79 epigenomes, 13 of which belong to different fetal tissues.

These datasets are available as Additional file 1.

In order to interpret the large number of genome regions detected by this procedure when they contain annotated genes, we retrieve these genes with Bedtools [26]. For that we use the latest annotation available for the genome assembly our epigenomic data is based on (GRCh37.p13), downloaded from the NCBI ftp site. Any gene overlapping, totally or partially, with the genome segment of interest is retrieved.

To assess whether a collection of genes (coming from a set of genome segments of interest) is representing a particular biological process or processes we performed a “functional enrichment analysis”. This standard analysis looks for functional keywords/terms overrepresented in the set of genes of interest respect to a background (the whole human genome in this case). For this we used the topGO R package [27], fed with the Gene Ontology (GO) [28] annotations for the human genes downloaded from the NCBI. We chose the Fisher’s exact test implemented in this package and subsequently applied Benjamini-Hochberg FDR correction for multiple testing for quantifying the overrepresentation of a given GO term in a set of genes.

We evaluate the effect of the number available epigenomes on the quality of the results obtained for the “brain” dataset. In this dataset it is possible to automatically detect the enriched GO terms unambiguously related to “brain” (i.e. those containing keywords such as “brain”, “neuron”, “nerve”, “axon”, ... and related), as a proxy for “positive” terms. For this dataset, we repeated the procedure described above randomly removing increasing proportions of the initial epigenomes (from 20 to 80%). Every removal was repeated 5 times and the results averaged. We quantify the performance of the method in the original dataset as well as in the reduced sets as the negative logarithm of the enrichment p -value for those “positive” GO biological processes. While this is far from being a real quantification of performance (e.g. many terms without these keywords, such as their parents, can also be positives) it allows us to compare its relative change in this down-sampling experiment.

Results

We applied the procedure described in Methods to the three classification schemas in order to detect the genomic regions associated to the phenomena behind the classifications. In the three cases, we scanned the whole

human genome with consecutive non-overlapping windows of length 5000 bp, applying the procedure described in Methods with a reduced vocabulary of 5 states, and took the segments with MI score of 0.3 or higher and a FDR-corrected p -value of 1E-4 or lower.

Brain samples

In the case of the brain tissues, it is expected that the genomic regions detected are in some way related to the function and development of this organ, since they present a characteristic epigenetic behavior in this organ different from that in others. We found 2991 regions in the genome with MI score 0.3 or higher and a FDR-corrected p -value of 1E-4 or lower. Chromosome 1 has the highest number of these regions (296). The lists of regions (sorted by the MI score) for each chromosome are available as Additional file 2. The lists include links to a genome browser to inspect the genomic neighbourhood of each window, as well as their epigenetic profiles in comparison with the group classification (brain vs. others).

These regions overlap with 1030 different genes according with the annotation procedure followed. Two hundred forty-six regions (8%) do not overlap with any gene. The enrichment analysis of Gene Ontology terms clearly shows that these 1030 genes are enriched in biological processes related to neuronal development and functioning, such as “nervous system development”, “neurogenesis”, “generation of neurons”, “neuron differentiation”, ... (Table 1). These genes are also enriched in Gene Ontology “cellular compartment” annotations (GO_CC) such as “neuron part”, “synapse”, “neuron projection”. The whole list of enriched terms is available in the Additional file 3, and a graphical summary of these, generated with REVIGO [29], in the Additional file 1: Figure S3.

The region with the highest MI score is 134,635,001–134,640,000 in chromosome 6. The epigenetic profile of this region shows that it tends to be in “active transcription” state in the brain samples while heterochromatized/repressed in the others (Fig. 2). This region matches part of an intron and the last exon of gene SGK1, a kinase that plays a role in neuron excitability by activating a number of ion channels. The next region in the list (Chr12:49255001–49,260,000) has a similar epigenetic profile and overlaps with intronic and exonic regions of gene RND1, a Rho GTPase involved in axon extension. The third region with highest score (Chr16:56310001–56,315,000) has another epigenetic profile: it is in “enhancer” state in brain tissues while heterochromatized in others. It matches an intronic region of gene GNAO1, a G protein repressor, whose defects are a cause of early-onset epileptic encephalopathy. The fourth hit (Chr3:171820001–171,525,000) has an

Table 1 Enriched processes in the three datasets. For the brain dataset, only the top of the list is shown. The whole lists with additional information (e.g. genes in each process), are available as Additional file 3 (*Continued*)

BRAIN			FETAL			CANCER		
GO.ID	Term	Fisher Test	GO.ID	Term	Fisher Test	GO.ID	Term	Fisher Test
GO:0044767	single-organism developmental process	9.80E-16						
GO:0032989	cellular component morphogenesis	2.80E-15						
GO:0000902	cell morphogenesis	2.90E-15						
GO:0051960	regulation of nervous system development	5.30E-15						
GO:0050803	regulation of synapse structure or activ...	8.80E-15						
GO:0007154	cell communication	1.60E-14						
GO:0007416	synapse assembly	2.10E-14						
GO:0050804	modulation of synaptic transmission	7.00E-13						
GO:0044707	single-multicellular organism process	8.10E-13						

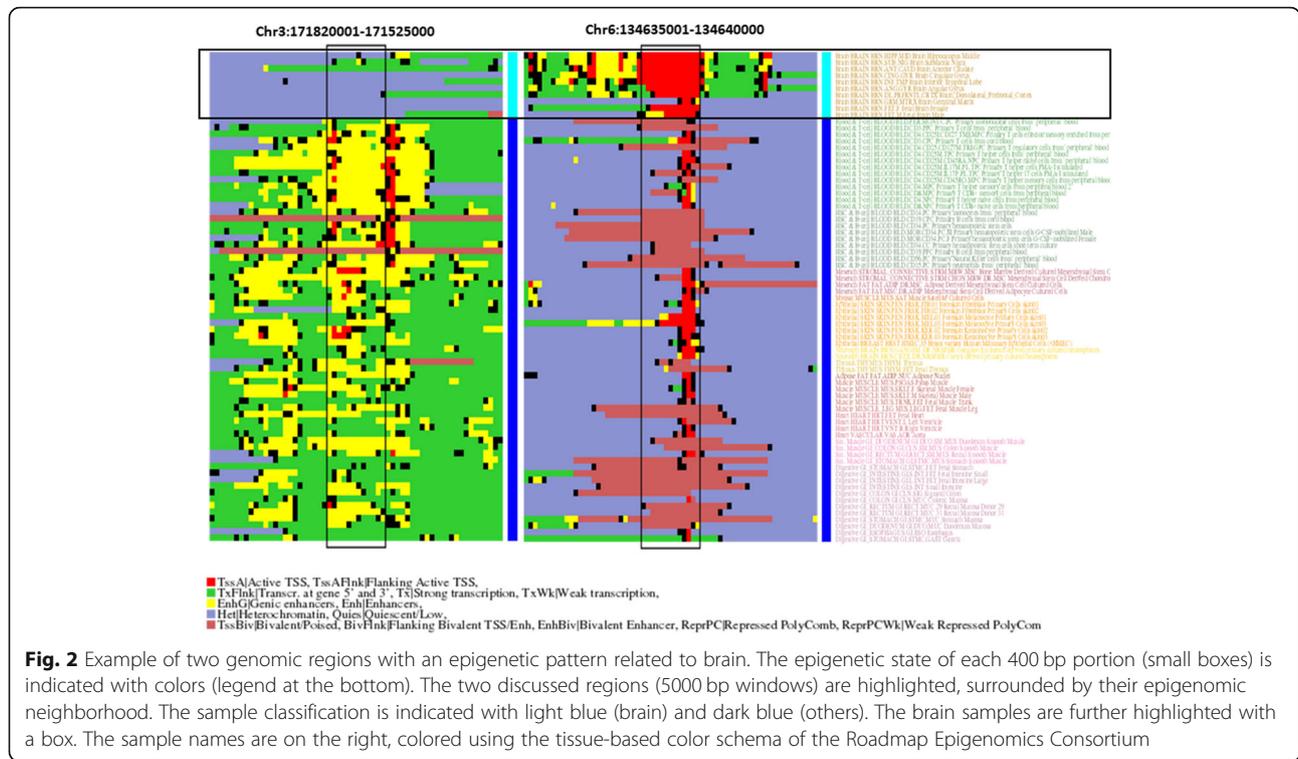


Fig. 2 Example of two genomic regions with an epigenetic pattern related to brain. The epigenetic state of each 400 bp portion (small boxes) is indicated with colors (legend at the bottom). The two discussed regions (5000 bp windows) are highlighted, surrounded by their epigenomic neighborhood. The sample classification is indicated with light blue (brain) and dark blue (others). The brain samples are further highlighted with a box. The sample names are on the right, colored using the tissue-based color schema of the Roadmap Epigenomics Consortium

opposite epigenetic profile: inactive/heterochromatin in most brain tissues, and enhancer/weak transcription in the rest (Fig. 2). This window corresponds to an intronic region of gene *FNDC3B*, a fibronectin type III containing protein without obvious relationship with the brain. Nevertheless, the expression of that gene in the brain (from the GTEx project [30]) is much lower than in any other tissue (Additional file 1: Figure S1), in agreement with its epigenetic profile and pointing to a relationship with this organ.

As shown in Fig. 5, the two most frequent patterns of epigenetic change between brain samples and those of other tissues in these 2991 windows is transcription (brain) / heterochromatin (others), followed by transcription (b) / bivalent chromatin (o). The third pattern is the opposite (heterochromatin (b) / transcription (o), and the fourth pattern is enhancer (b) / heterochromatin (o). We performed the GO enrichment analysis described above for the windows within these different sets independently, to get insight into the molecular processes associated to these different epigenetic changes. In the following, we focus on the GO terms differentially enriched in these subsets (respect to those enriched in the whole set commented above). The two first patterns, which involve specific activation of transcription in brain samples, are related to neuron development and function (biological processes) and synapsis and ion channels (cellular compartments). The opposite pattern (silencing in brain) is related to “apoptosis”, “cell death” and

similar terms (In GO:CC, the term differentially enriched for this pattern is “intracellular”). This points to a scenario where processes related to neuron development are “activated” in brain while those related to cell removal/turnover are “silenced”. For the fourth pattern, the enriched terms are related to nucleotide metabolic processes. The detailed results of this subset enrichment analysis are available as Additional file 4.

The accumulation of these high-MI 5000 bp windows in certain regions can also point to interesting sites. That is, even if none of these windows is at the top of the list, as the four we discussed above, the fact that they concentrate in a particular region of the genome could be informative. Probably that region would show up scanning with a larger window. The gene with the largest number of high-MI windows is *CSMD2*. Fifty out of the 2991 windows match this long gene. This gene is involved in the control of complement, and defects on it have been associated with schizophrenia. The next gene is *CTNND2* (49 windows). This gene encodes an adhesive junction associated protein involved in brain and eye development, and it resides in a region of chromosome 5 deleted in the *Cri du Chat* syndrome, characterized by intellectual disability and microcephaly. The other genes with high concentration of informative windows are also related to brain in one way or another (data not shown).

We used this “brain” dataset to evaluate the effect of the number of available epigenomes on the performance

of the method, since the “correctness” of the final set of GO:BP enriched terms is easier to assess automatically in this dataset (see Methods). Additional file 1: Figure S2 shows the variation of performance as we remove increasing proportions of the input epigenomes. It can be seen that the results are quite stable until we remove a large proportion of the epigenomes (60%), a point where it drops sharply. When 80% of the epigenomes are removed, we either ended up without any brain sample in the input set of epigenomes (i.e. no results are generated) or the results are bad and no brain-related GO:BP terms showed up as enriched.

Fetal samples

We found 224 regions in the genome associated to this phenomenon with the procedure described above (Additional file 2), with chromosome 7 containing the largest number of them (38), followed by chromosome X (22).

These windows overlap with 130 annotated genes. 20% of them (44 regions) do not overlap with any gene. These genes are enriched in biological functions related to organism development and signal transduction, such as “regulation of signal transduction”, “neurogenesis”, “ossification” and morphogenesis of different organs (Table 1). There are no “cellular compartment” (CC) terms enriched with the cutoffs used. The whole list of enriched terms is available in the Additional file 3, and a graphical summary of these, generated with REVIGO [29], in the Additional file 1: Figure S4.

The region with the highest score is Chr7:23385001–23,390,000). Its epigenetic profile shows that it is being

transcribed in the fetal samples while heterochromatinized in most of the others (Fig. 3). This region overlaps with intronic and exonic regions of gene IGF2BP3, an insulin-like growth factor 2 mRNA binding protein that repress the translation of that protein during late development. The second window with highest score is Chr2:11530001–11,535,000, which partially overlaps with the long non-coding RNA LINC00570, of unknown function. This region has the “enhancer” state in fetal samples and a mixture of heterochromatin and transcription in the others. The next region in the list, in chromosome 8, show a similar epigenetic pattern and there are not annotated genes on it. The fourth high scoring window is again in chromosome 7 (130655001–130,660,000) and has an epigenetic behavior opposed to the first one discussed above: it is heterochromatinized in the fetal samples while transcribing in most of the others (Fig. 3). This window overlaps with a long non-coding RNA induced by p53 (LOC378805). The whole list of high scoring windows, split by chromosomes is available as Additional file 2.

The first and second genes commented above (IGF2BP3 and LINC00570) also are those with the highest concentration of high-scoring windows (11 and 21 respectively), followed by IGF2BP1 (functionally similar to IGF2BP3) with 7 windows.

Regarding the patterns of epigenetic change between fetal and adult tissues in these windows, Fig. 5 shows a plainer distribution than in the other two datasets, indicating that the preference for certain transitions is not so clear. Nevertheless, the three most frequent transitions are

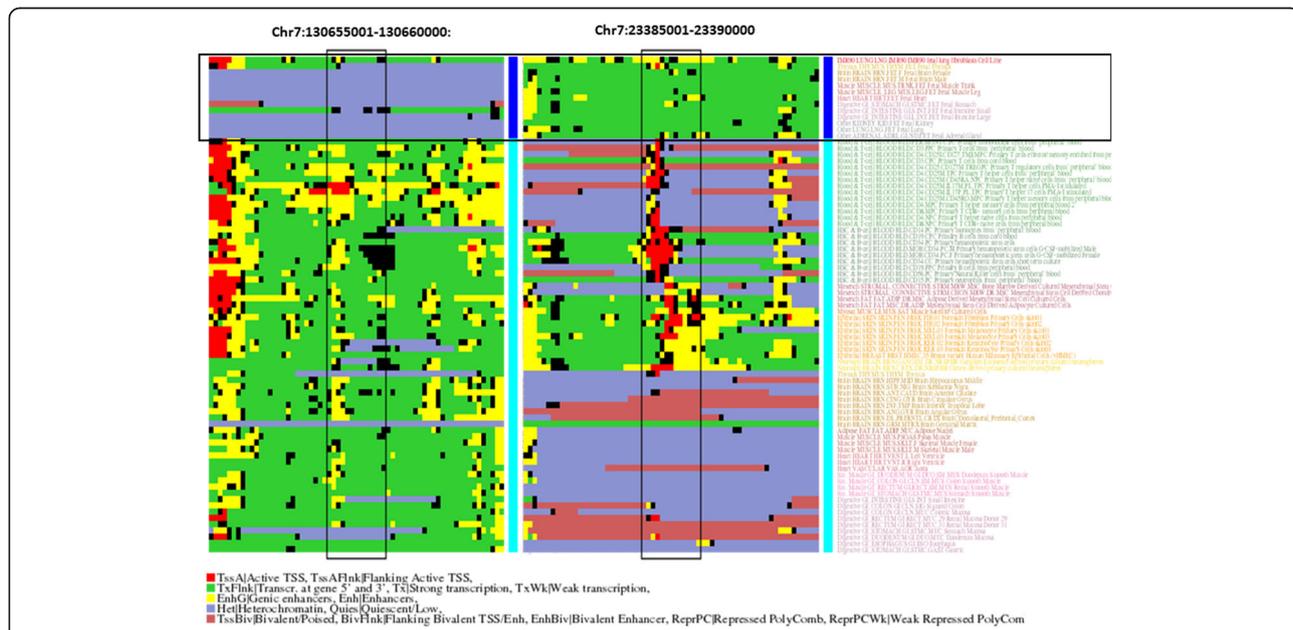


Fig. 3 Example of two genomic regions with an epigenetic pattern related to fetal samples. Same representation as in Fig. 2 for fetal samples (dark blue) vs. others (light blue). By the Roadmap coloring of the samples it is evident that the “fetal” set contains now a variety of (fetal) tissues/organs

those from transcription, enhancer and bivalent states in fetal to heterochromatin in adult, followed by two patterns representing a somehow opposite transition: transcription in adult vs. heterochromatin/bivalent in fetal. The GO terms associated to the windows within these five subsets are related, respectively, to *none*, extracellular matrix, neuron biogenesis/cell cycle, nucleotide metabolic processes and ossification/immune response (Additional file 4).

Cancer samples

In this case we detected 345 regions in the human genome, with chromosomes 1 and 2 inclosing the largest number of them (33 each) (Additional file 2).

These regions overlap with 185 annotated genes. Ninety-six regions (28%) do not overlap with any gene. This set of genes is enriched in biological processes clearly related to cancer, such as “cell cycle”, “mitotic cell cycle”, “regulation of transcription involved in cell fate” and “microtubule-based movement” (Table 1). The only enriched GO_CC term is “kinesin complex”. As with the other datasets, the whole list of enriched terms can be found in the Additional file 3, and their REVIGO graphical summary in Additional file 1: Figure S5.

There are five regions with the highest score (0.994) and a *p*-value of 0.0, two of them in chromosome 7, and the others in chromosomes 6 and 4. All overlap with genes known or potentially involved in cancer: JAZF1 (transcriptional repressor for which chromosomal aberrations involving it are associated with stromal tumors), PLXN4A (involved in cytoskeleton remodeling), SCML4 (involved in a complex required to maintain the transcriptionally repressive state of homeotic genes throughout development) and RBPJ (a transcriptional regulator important in the Notch signaling pathway), respectively.

Figure 4 shows the epigenetic profiles for two of these regions: those in chromosomes 7 and 6. In both cases the epigenetic patterns would be in agreement with the roles of the genes coded there. For example, the region

in chromosome 7 (gene JAZF1) is active in healthy tissues, hence producing the transcriptional repressor, while in cancer tissues it is heterochromatinized and, consequently, the repressor is not being transcribed.

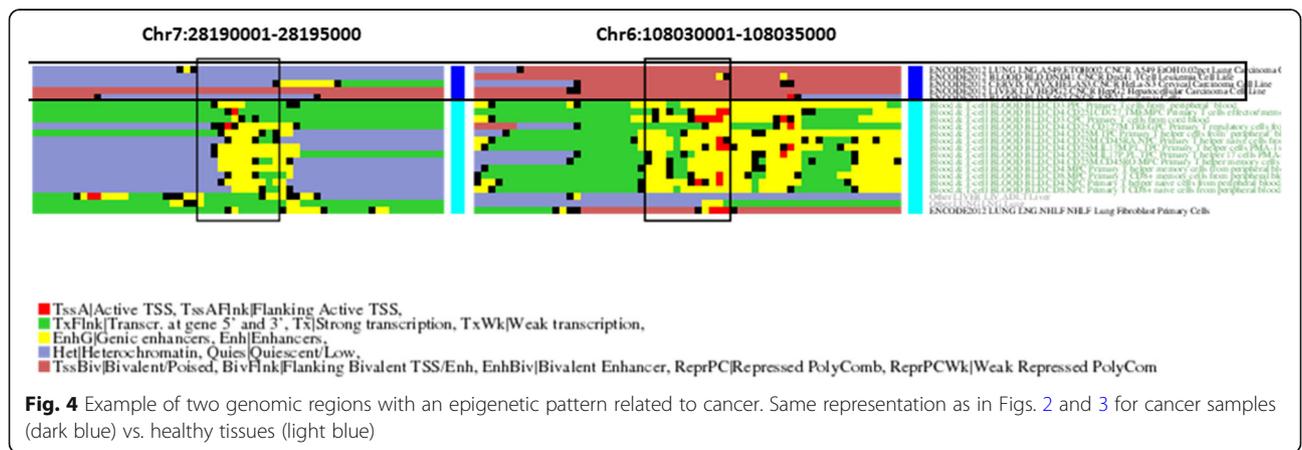
In spite of these particular examples, in the cancer-related genomic segments, globally the most frequent patterns of epigenetic change between cancer and healthy tissues are transcription (cancer) / heterochromatin (healthy) and bivalent (c) / heterochromatin (h) (Fig. 5). The first pattern is differentially enriched in terms related to neuron action potential, while there are no differentially enriched terms for the second (Additional file 4).

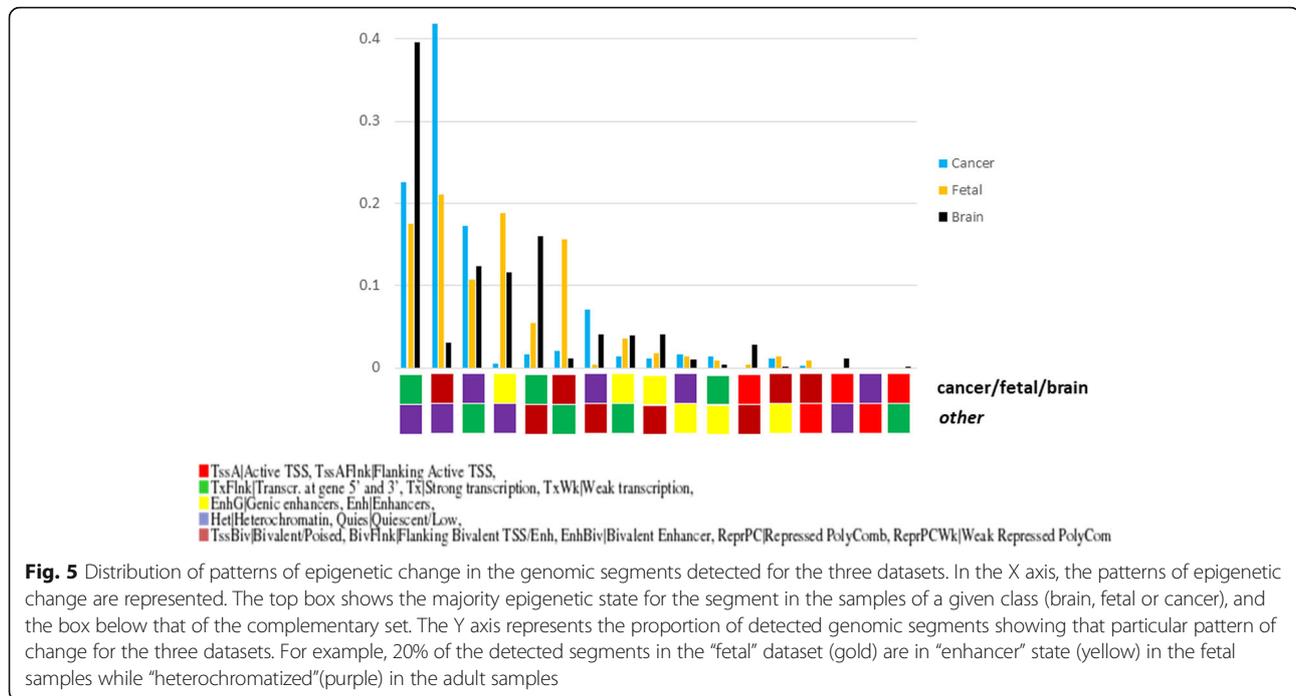
As with the other examples, the concentration of 5000 bp windows in a given gene is also informative. The gene with the highest accumulation of windows is LYST (14 windows), a lysosomal trafficking regulator whose relationship with cancer is not evident. The next one (10 windows) is KIF4A, a kinesin involved in maintaining chromosome integrity during mitosis and organizing the mitotic spindle prior to cytokinesis.

Discussion

While the DNA sequence of the human genome is virtually the same across all developmental stages, cell types and tissues, its epigenetic state varies, largely contributing to this cellular phenotypic diversity. Consequently, the massive epigenomic data that are being generated are expected to enclose a lot of new information on chromatin dynamics during development, cell differentiation and disease. Resembling what happened with genomic data, many computational approaches are being developed to extract that information from the massive epigenomic datasets.

The approach we present here allows to automatically detect functional regions in the genome, as those with patterns of inter-sample epigenetic variations correlated with an external classification of these samples in an arbitrary number of groups. Not only the regions are





detected but it is also possible to get insight into their role by looking into more detail on their particular variation of epigenetic patterns (e.g. active in a group of samples while inactive in others, enhancer, etc.).

Even if it is not an exhaustive test, the results of the “downsampling” experiment with the brain dataset seems to indicate that the method is quite robust to missing data (i.e. lack of epigenomes). The results with the cancer dataset, consisting only in 21 epigenomes, points into the same direction. In any case, we expect the epigenomic data to increase substantially in the future.

The same set of samples can be used to extract different types of functional regions as long as they can be classified in different ways, as we showed here with three orthogonal classifications of the same Roadmap Consortium epigenomes. In this sense the method can be regarded as “supervised”, as the sample classification is provided by the user. This is a fundamental difference with another approach for detecting genome regions with a particular epigenetic pattern, ChromDet [22]. ChromDet is “unsupervised”, in the sense that it detects the regions better reflecting the main tendencies in the set of epigenomes given as input. Another difference is that, due to its computational cost, the method presented here can be applied to large sets of epigenomes and to exhaustively scan the whole genome in detail (i.e. with small windows). In spite of being suitable for scanning whole genomes, epigenetic information for the whole human genome is not strictly required for this approach, and it could work with epigenetic data for a

limited region (across a set of samples), while ChromDet requires whole (or very large) epigenomes for extracting their main tendencies. The recently developed ChromSwitch method [19] is also intended to locate regions where certain epigenetic marks change drastically between two sets of samples and, as in our case, it can also work with epigenomic information for a limited region. Apart from using a totally different methodology, the two main differences with the approach described here are its limitation to two groups of samples and the fact that it uses the raw quantitative epigenetic marks as input, instead of a vocabulary of states. Consequently, all these approaches complement each other in the search for chromosomal regions with informative epigenetic profiles.

Although we focused our large scale (enrichment-based) evaluation on regions containing genes, the method is intended to scan the whole genome and detect functional non-coding regions such as enhancers, regulatory regions, etc. Indeed, many of the regions we detected do not overlap with genes. A more detailed inspection of them will be required to assess whether they could be functionally or not (e.g. presence of binding sites for transcription factors, functional RNAs, etc.)

The sliding-window approach allows scanning the genome with different window sizes and moving them at different steps so that, for example, once a large segment of interest is found, smaller windows could be used to scan it in more detail.

As with many genomic studies, this approach still has the problem of distinguishing correlation from causation

[31]. It is difficult to assess whether the epigenetic behavior of the genomic regions we are detecting related to a given phenomenon is a cause of that phenomenon or just a consequence/marker of it. Further inspection of the actual epigenetic patterns could shed light into this. In the three datasets we found examples of windows with both between-classes transitions: transcription > heterochromatin and the other way around (heterochromatin > transcription), and in some cases these transitions make sense in the light of the molecular processes associated to the genes of the corresponding windows. In case we can detect regions whose epigenetic variation is causative for the studied phenomenon/disease, these results would be useful for a field that is receiving increasing attention, “epigenome editing” [31, 32]: i.e. modify epigenetics by changing the DNA motifs responsible for it (e.g. with CRISPR/Cas9) or by manipulating the editing enzymes. In the second case, when the detected epigenetic pattern is not causative of the phenomenon but just a reflection of it, it could be used as a marker to detect/monitor it.

Conclusions

The simple and scalable methodology presented here is able to detect functional genomic regions related to a given phenomenon using epigenomic information.

Paralleling what happened in genomics, we foresee a large accumulation of epigenomes in the future, even at the single-cell level [33]. So, methods such as that presented here, for mining that massive information and translating it into knowledge, would be valuable.

Additional files

Additional file 1: Additional figures and data. Figures S1, S2, S3, S4 and S5. Epigenome classifications for the three experiments. Detailed description of all epigenomes used. (DOCX 425 kb)

Additional file 2: Full list of windows in all chromosomes for the three experiments (in plain text and html format). The lists include the detailed epigenetic profiles for the windows as well as links to inspect these in a genome browser. (ZIP 973 kb)

Additional file 3: Full lists of GO enriched terms (BP: biological processes, and CC: cellular compartment) for the three experiments, including the genes overlapping with the detected windows within each GO term. (ZIP 33 kb)

Additional file 4: Full lists of GO enriched terms (BP: biological processes, and CC: cellular compartment) for the subsets of windows showing particular patterns of epigenetic change (Fig. 5). (ZIP 72 kb)

Acknowledgements

We sincerely thank Jesus de Celis, David Juan (UPF), Alfonso Valencia (BSC), Juan AG Ranea (U. Málaga) and the members of the Computational Systems Biology Group (CNB-CSIC) for insightful discussions. We also want to acknowledge Carlos O. Sanchez-Sorzano (CNB Statistical Support Facility) for his advice on some aspects of this work.

Funding

This work was partially funded by the Spanish Ministry for Economy and Competitiveness [SAF2016–78041-C2–2-R to F.P.]

Availability of data and materials

- All software developed for carrying out the preset study is available from the corresponding author on reasonable request.
- All data generated or analysed during this study are included in this published article [and its supplementary information files].

Authors' contributions

FP conceived the approach and performed preliminary tests. AGM performed additional tests and the GO enrichment analysis. JCO performed additional tests and the qualitative interpretation of the results. FP wrote the first draft of the manuscript. All authors help revising/refining the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 13 July 2018 Accepted: 20 November 2018

Published online: 28 November 2018

References

1. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet.* 2003; 33(Suppl):245–54.
2. Friedman N, Rando OJ. Epigenomics and the structure of the living genome. *Genome Res.* 2015;25(10):1482–90.
3. Rivera Chloe M, Ren B. Mapping Human Epigenomes. *Cell.* 2013;155(1):39–55.
4. Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL, et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell.* 2013;152(3): 642–54.
5. Atlasi Y, Stunnenberg HG. The interplay of epigenetic marks during stem cell differentiation and development. *Nat Rev Genet.* 2017;18(11):643–58.
6. Berdasco M, Esteller M. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Dev Cell.* 2010;19(5):698–711.
7. Jones PA, Issa JP, Baylin S. Targeting the cancer epigenome for therapy. *Nat Rev Genet.* 2016;17(10):630–41.
8. Rozek LS, Dolinoy DC, Sartor MA, Omenn GS. Epigenetics: relevance and implications for public health. *Annu Rev Public Health.* 2014;35:105–22.
9. Consortium EP. A User's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 2011;9(4):e1001046.
10. Fernández JM, de la Torre V, Richardson D, Royo R, Puiggròs M, Moncunill V, Fragkogianni S, Clarke L, Flicek P, Rico D, et al. The BLUEPRINT Data Analysis Portal. *Cell Syst.* 2016;3(5):491–495.e495.
11. Consortium RE, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518:317–30.
12. Belyaeva A, Venkatachalapathy S, Nagarajan M, Shivashankar GV, Uhler C. Network analysis identifies chromosome intermingling regions as regulatory hotspots for transcription. *Proc Natl Acad Sci U S A.* 2017;114(52):13714–9.
13. Mourad R, Cuvier O. Predicting the spatial organization of chromosomes using epigenetic data. *Genome Biol.* 2015;16:182.
14. Zhu Y, Chen Z, Zhang K, Wang M, Medovoy D, Whitaker JW, Ding B, Li N, Zheng L, Wang W. Constructing 3D interaction maps from 1D epigenomes. *Nat Commun.* 2016;7:10812.
15. Li W, Wang M, Sun J, Wang Y, Jiang R. Gene co-opening network deciphers gene functional relationships. *Mol Biosyst.* 2017;13(11):2428–39.

16. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Varnai C, Thiecke MJ, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*. 2016;167(5):1369–1384.e1319.
17. Burgess DJ. Epigenomics: deciphering non-coding variation with 3D epigenomics. *Nat Rev Genet*. 2016;18(1):4.
18. Rakyán VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*. 2011;12:529.
19. Jessa S, Kleinman CL. Chromswitch: a flexible method to detect chromatin state switches. *Bioinformatics*. 2018;34(13):2286–8.
20. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*. 2017;12(12):2478–92.
21. Hoffman MM, Buske OJ, Wang J, Weng Z, Billes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods*. 2012;9(5):473–6.
22. Carrillo-de-Santa-Pau E, Juan D, Pancaldi V, Wera F, Martín-Subero I, Rico D, Valencia A, on behalf of the BLUEPRINT C. Automatic identification of informative regions with epigenomic changes associated to hematopoiesis. *Nucleic Acids Res*. 2017;45(16):9244–59.
23. Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. *Nat Rev Genet*. 2013;14(4):249–61.
24. Chakraborty A, Chakrabarti S. A survey on prediction of specificity-determining sites in proteins. *Brief Bioinform*. 2015;16(1):71–88.
25. Chagoyan M, Garcia-Martin JA, Pazos F. Practical analysis of specificity-determining residues in protein families. *Brief Bioinform*. 2016;17(2):255–61.
26. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.
27. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology; 2016.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25–9.
29. Supek F, Bosnjak M, Skunca N, Smuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*. 2011;6(7):e21800.
30. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. The genotype-tissue expression (GTEx) project. *Nat Genet*. 2013;45:580.
31. Stricker SH, Koflerle A, Beck S. From profiles to function in epigenomics. *Nat Rev Genet*. 2017;18(1):51–66.
32. Mendenhall EM, Williamson KE, Reyón D, Zou JY, Ram O, Joung JK, Bernstein BE. Locus-specific editing of histone modifications at endogenous enhancers. *Nat Biotechnol*. 2013;31(12):1133–6.
33. Schwartzman O, Tanay A. Single-cell epigenomics: techniques and emerging applications. *Nat Rev Genet*. 2015;16(12):716–26.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

