bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Evaluating the Stability and Flexibility of DNA Methylation Patterns from Stem to Differentiated Cells

Minseung Choi[1,3,◕,*], Diane P. Genereux[2,4,◕,*], Jamie Goodson[5], Haneen Al-Azzawi[6], Shannon Q. Allain[1], Stan Palasek[7], Carol B. Ware[8,9], Chris Cavanaugh[8,9], Daniel G. Miller[9,10], Winslow C. Johnson[11], Kevin D. Sinclair[6], Reinhard Stöger[6], Charles D. Laird[1,12,*]

1 Department of Biology, University of Washington, Seattle WA 98195
2 Department of Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester MA 01605
3 Department of Computer Science, Princeton University, Princeton NJ 08544
4 Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge MA 02142
5 Department of Pathology, University of Washington School of Medicine
6 School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough, Leicestershire LE12 5RD, UK
7 Department of Mathematics, Princeton University
8 Department of Comparative Medicine, University of Washington School of Medicine
9 Institute for Stem Cell and Regenerative Medicine, University of Washington
10 Department of Pediatrics, University of Washington School of Medicine
11 Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260
12 Department of Genome Sciences, University of Washington

◕These authors contributed equally to this work.

* mschoi@princeton.edu, genereux@broadinstitute.org, cdlaird@uw.edu

Abstract

DNA methylation has been studied extensively in many developmental systems. Little attention, however, has been given to identifying methylation features that distinguish loci whose patterns are in flux in a given cell lineage from those whose patterns are stable. Here, we develop a new metric, the Ratio of Concordance Preference (RCP), to quantify and compare epigenetic flexibility and stability across loci, cell types, and developmental stages, without assuming any specific biochemical mechanisms. We apply RCP to double-stranded DNA methylation data from human and murine cells and conclude that: (i) preference for concordant DNA methylation is reduced but not eliminated in stem relative to differentiated cells; (ii) cellular differentiation is characterized by increasing preference for concordant methylation states; and (iii) while concordance preference remains substantial through embryonic totipotency and early stages of pluripotency, primordial germ cells initially have nearly no preference for concordance, perhaps reflecting the high level of epigenetic flexibility *en route* to production of gametes. The mechanism-free nature of RCP will enable

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

comparison of DNA methylation systems not only across cell types and developmental stages, but also across organisms whose methylation machineries are not well understood or may differ significantly.

Author Summary

In storing and transmitting epigenetic information, organisms must balance the need to maintain information about past conditions with the capacity to respond to current and future information from their environments. Methylation states can be transmitted with variable levels of fidelity from parent to daughter strand: high fidelity confers strong pattern matching between the strands of individual DNA molecules and thus pattern stability over rounds of DNA replication; lower fidelity confers reduced pattern matching, and thus greater flexibility. We are interested in the strategies that various cell types, organisms, and species use to achieve balance between flexibility and stability. Here, we present a new conceptual framework, the Ratio of Concordance Preference (RCP), that can quantify the flexibility and stability of the system that gave rise to a given set of DNA methylation patterns. We confirmed previous observations that differentiated cells operate with high DNA methylation stability. In contrast, increased flexibility was inferred for a wide variety of stem cells, although substantial levels of stability remained. One exception was the near-complete, albeit transient, potential loss of stability in the precursors of eggs and sperm. Broader application of our RCP framework will permit comparison of epigenetic-information systems across cells, developmental stages, and organisms whose methylation machineries differ substantially or are not yet well understood.

Introduction

Organismal development is characterized by a shift from the phenotypic flexibility of embryonic cells to the canalized identities of the lineages that lead to differentiated cells. To achieve stable gene-regulatory states in terminally differentiated cells, organisms ranging from archaea to humans use a variety of epigenetic mechanisms, including DNA methylation. Perturbation of the state of DNA methylation at various loci in differentiated cells is associated with several human cancers [1–3]; in turn, restoring epigenetic flexibility of some loci has proven challenging in efforts to produce induced pluripotent stem (iPS) cells [4]. Together, these findings highlight the importance of shifts between epigenetic flexibility and stability in establishing cellular identity.

There exists an extensive literature documenting changes in single-locus and genome-wide methylation frequencies at various stages of development [5, 6]. Most genomic regions in primordial germ cells (PGCs), for example, are known to undergo dramatic and rapid shifts in DNA methylation frequency [7]. Recently, it has become clear that active demethylation can occur in mammalian stem and iPS cells [8], highlighting the potential for both gain and loss of cytosine methylation to impact the overall methylation frequency and stability of a given genomic region during development.

High concordance of methylation in differentiated cells, with matching states for parent and daughter DNA strands at individual CpG/CpG dyads, is considered to be a

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

hallmark of conservative epigenetic processes [9–13]. For earlier stages of development, the extent of concordance is far less clear. For example, do methylation patterns in dividing embryonic stem cells arise entirely by random placement of methyl groups, or is concordance favored to some degree?

Recent papers have begun to address these issues [7, 14–18]. Shipony *et al.* [16] analyzed single-stranded DNA methylation patterns in populations of cultured cells established from single founder cells. Under this approach, the degree of stability was inferred from the extent of congruence among patterns collected from cultured descendants. The observation of substantial diversity among patterns in the descendant cells led Shipony *et al.* [16] to conclude that the bulk of methylation in human embryonic stem (ES) and induced pluripotent stem (iPS) cells arises through "dynamic" — or non-conservative — DNA methylation processes rather than through the "static" — or conservative — processes that were emphasized in earlier studies [10, 11, 19]. Using hairpin-bisulfite PCR [13], a method that yields double-stranded DNA methylation patterns, other studies concluded that dynamic processes contribute substantially to DNA methylation in cultured mouse ES cells, but perhaps not to the exclusion of the conservative processes that dominate at many loci in adult differentiated cells [7, 14, 15, 17, 18].

To fully understand the balance between conservative and non-conservative methylation processes, it is necessary to quantify the extent to which the arrangement of methylation in a given set of patterns deviates from the null assumption of random placement. To assess and visualize such deviations, we here introduce a new metric, the Ratio of Concordance Preference (RCP), which utilizes double-stranded methylation patterns. Here, as previously, we use the term *double-stranded DNA methylation pattern* to refer to a pattern that provides information on the methylation patterns on both the top and the bottom strands of an individual double-stranded DNA molecule. These double-stranded patterns provide information on the extent of matching between methylation states on parent and daughter strands, which are separated by exactly one round of DNA replication. RCP requires no assumptions about the enzymatic mechanisms of methylation and demethylation, and so enables comparison across diverse species and developmental stages. In a recent review, Jeltsch and Jurkowska [20] emphasized the balance of methylating and demethylating processes — rather than the propagation of specific individual methylation patterns — as the primary determinant of the overall set of methylation patterns present in a given cellular population at a given time. In this framework, RCP can be thought of as a metric for quantifying the extent to which the patterns produced by a given system of methylating and demethylating processes deviate from the sets of patterns expected if methyl groups are placed entirely at random.

In parameterizing RCP, we use the term "conservative", in lieu of "static" as used previously [16], to describe processes that preferentially establish concordant as opposed to discordant methylation states. We consider non-conservative processes, described previously as "dynamic" [16], as having one of two forms: "random" processes that add or remove methyl groups with equal preference for concordance and for discordance, and "dispersive" processes that preferentially establish discordant methylation states.

We validate our RCP framework by confirming its ability to identify systems with near-dominance and near-absence of conservative processes, as well as systems on the continuum between these extremes. We apply this new method to our authenticated, double-stranded DNA methylation patterns, both published and previously unpublished, collected from human and murine cells. To expand the data available for our initial

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

RCP analysis, we also examine patterns from three recent publications that utilized hairpin-bisulfite PCR of murine DNA [14, 15, 17]. To improve our understanding of the processes by which stem cells are reprogrammed to give rise to differentiated cells, we ask: (*i*) how strong the preferences are for concordant DNA methylation states in cultured and embryonic stem cells; (*ii*) whether this preference changes as stem cells shift between undifferentiated and differentiated states; and (*iii*) how preference changes in cells during embryonic development.

Results and Discussion

Defining the Ratio of Concordance Preference for All Possible Methylation Configurations

We have developed the Ratio of Concordance Preference (RCP) to assess the strategy of binary information transfer, with focus on the degree to which exact information is conserved. We apply our RCP framework to DNA methylation in mammalian cells. Our goal is to infer whether and how much the system of processes that established a given set of methylation patterns prefers concordant to discordant methylation states. This general formulation is free of assumptions about the molecular mechanisms whereby methylation is added to and removed from DNA.

In our data from double-stranded DNA molecules from human and mouse, methylation occurs principally at the CpG motif. This symmetric motif may be written as CpG/CpG, here termed "CpG dyad". CpG dyads have opportunities for methylation on both strands. The methylation state of a dyad is thus of one of three forms: fully methylated, at frequency M, with methylated cytosines on both strands; hemimethylated, at frequency H, with a methylated cytosine on only one of the two strands; and unmethylated, at frequency U, with neither cytosine methylated. The RCP framework can also be extended to non-CpG methylation at symmetric nucleotide motifs.

To infer concordance preference for sets of double-stranded methylation patterns, we use the overall frequency of methylation, m, and the frequency of unmethylated dyads, U, of each data set. Because m is derived from the three dyad frequencies, the pair (m, U) encompasses the full information available from the implicit dyad frequencies, M and H. We evaluate the extent of deviation from expectations under a random model in which the system has no preference for either concordant or discordant placement of methyl groups, using RCP, defined as:

$$\mathrm{RCP} = \frac{\sqrt{U(U+2m-1)}}{1-U-m} \tag{1}$$

We note that the expectations of a random model, for which RCP = 1, are met both by truly random placement of groups, and by equal contributions from processes operating with strong preference for concordance and processes operating with strong preference for discordance. Under this random model, the frequency of unmethylated dyads is given by $U = (1-m)^2$, leading to dyad frequencies as expected under the binomial distribution (Fig 1a-b dashed curve; Fig 1c). A system for which methylation processes are dominated by the activity of the mammalian Dnmt3s, which add methyl

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

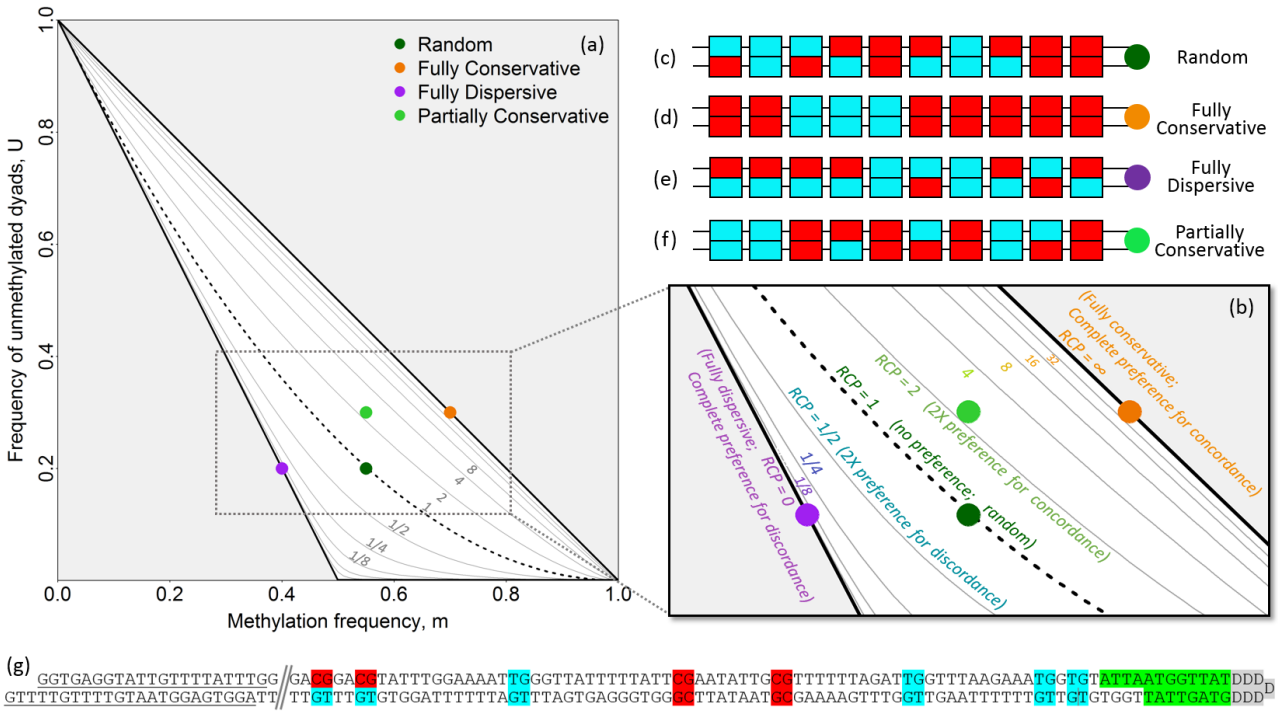


Figure 1. Characterizing Methylation Systems Using Double-Stranded DNA Methylation Patterns. (a) Frequencies of methylated cytosines (m) and unmethylated dyads (U) locate each data set on the continuum from complete preference for concordance to complete preference for discordance. Ratio of Concordance Preference (RCP) is indicated for each contour line. **(b)** Expanded view. For this schematic, individual double-stranded methylation patterns **(c-f)** are used to illustrate different methylation configurations that lie along this continuum. Individual patterns, with methylated and unmethylated cytosines indicated in red and blue, respectively, can reflect **(c)** a random methylation system; **(d)** a fully conservative system, with complete preference for generating concordant dyads; **(e)** a fully dispersive system, with complete preference for generating discordant dyads; or **(f)** a partially conservative system, with more concordant dyads than expected under random processes, but fewer than expected under fully conservative processes. Systems can also exhibit partial preference for dispersive placement of methyl groups. **(g)** A representative partial double-stranded DNA methylation pattern collected using hairpin-bisulfite PCR. The experiment-specific batchstamp is shown in green, and can be used to monitor for PCR contamination; the molecule-specific barcode shown in gray, generalized as "DDDDDDD", can be used to identify redundant sequences. The batchstamp and barcode are encoded on the hairpin oligonucleotide used to join the top and bottom strands. Primer-binding sites are underlined at the left end of the molecule.

groups principally without regard to the methylation state of the other strand [21], is expected to behave largely in accordance with the random expectation.

One set of deviations from the random expectation is characterized by preference for concordant placement of methyl groups, such that the two classes of concordant dyads — fully methylated and fully unmethylated — are more frequent than expected under the random model. This situation occurs under conservative systems of methylation where strong contributions from maintenance-like processes, such as the activity of Dnmt1 in mammals [11, 13, 22], lead to high frequencies of concordant dyads. In the extreme form of this deviation from random, methyl groups are observed only in fully methylated dyads (Fig 1d), such that unmethylated dyads occur at frequency $U = 1 - m$ (upper diagonal line in Fig 1a-b).

The other set of possible deviations from random is characterized by preference for discordant placement of methyl groups, leading to an overabundance of hemimethylated dyads. This situation occurs under dispersive systems of methylation such as those that yield transient hemimethylation in the interval after DNA replication and prior to

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

daughter-strand methylation, and perhaps in genomic regions undergoing demethylation during periods of epigenetic transition. When methylation is maximally dispersive, and methylation frequency m is less than 0.5, all dyads with methylation will be hemimethylated (Fig 1e), such that $U = 1 - 2m$ (lower diagonal line in Fig 1a-b); when m is greater than 0.5, not all methyl groups can be accommodated in hemimethylated dyads, and so a combination of hemimethylated and fully methylated dyads — but no unmethylated dyads — is expected (lower horizontal line in Fig 1a).

The two extreme deviations from random form the boundaries of the comprehensive space of possible configurations of methylation at symmetric motifs (Fig 1a). Sets of double-stranded methylation patterns are expected to fall on the continuum between the extreme expectations (Fig 1b), and can be located within this space to characterize the processes that gave rise to a given data set, ranging from conservative to dispersive.

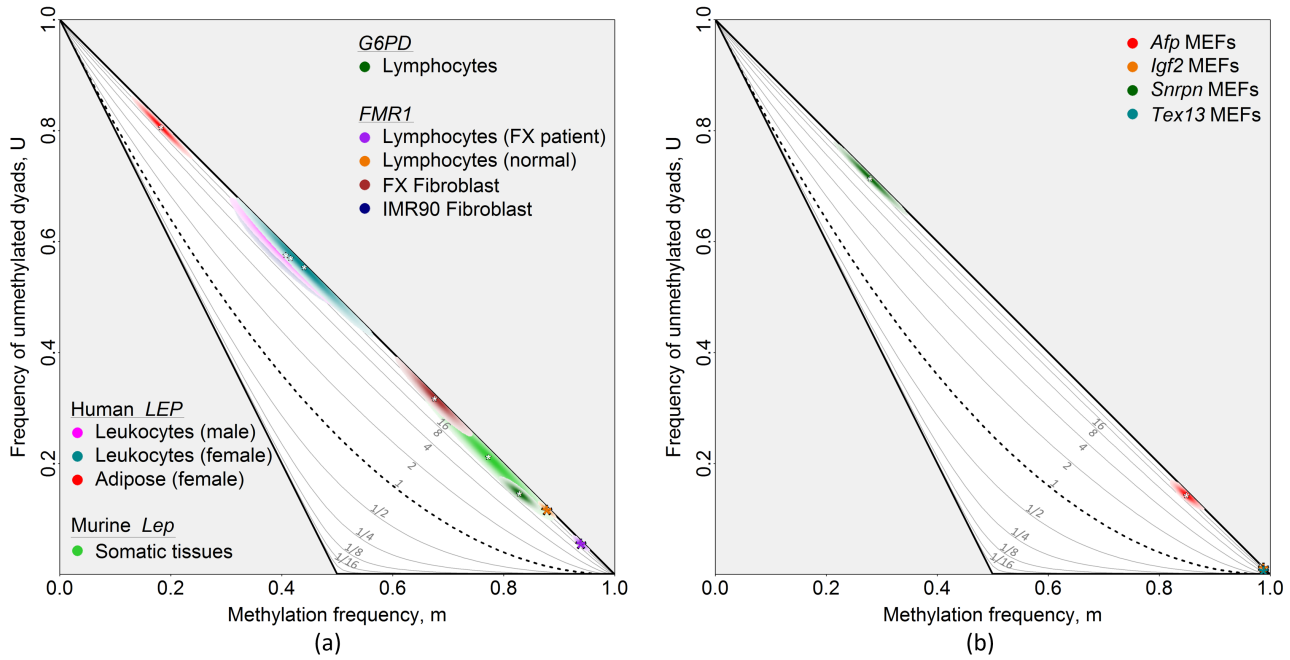
As noted above, the "Ratio of Concordance Preference" (RCP), for which the mathematical derivation is given in Supporting Information (S1 Text), measures whether and how severely the dyad frequencies of a given set of double-stranded DNA methylation patterns deviates from the assumption that methyl groups are arranged at random on DNA, with fully methylated, hemimethylated, and unmethylated dyads present at the frequencies expected under the binomial distribution. This expectation is analogous to the distribution of genotype frequencies at a two-allele locus in a population that is at Hardy-Weinberg equilibrium [23, 24]: RCP^2 is equal to $4MU/H^2$, which is expected to equal 1 at Hardy-Weinberg equilibrium.

A system with an RCP value of 1 has no preference for either concordance or discordance of methylation, and follows the random model (Fig 1c). An RCP value of 2 indicates two-fold preference for concordance, while an RCP of $\frac{1}{2}$ indicates two-fold preference for discordance. RCP approaches infinity for systems that have complete preference for concordant dyads. At the other extreme, RCP approaches 0 (i.e., $\frac{1}{\infty}$) for systems that have complete preference for discordant dyads. For the examples analyzed here, data for different loci and cells range from complete concordance to near-random positions on the RCP spectrum. Complete discordance is found as a transient condition of adenine methylation at the *ori* locus in *Escherichia coli*, and serves to regulate the timing of reinitiation of DNA synthesis [25, 26]. Dispersive methylation is also expected to exist ephemerally in the instant between DNA replication and methylation. During this instant, all methyl groups present are expected to be on the parent strand, as groups have yet to be added to the daughter. Thus, a broad spectrum of concordance preference can exist in organisms, and can be quantified and evaluated by RCP.

For large and intermediate-size data sets, the resolution of RCP is high across the range of possible methylation frequencies, although the resolution declines as m closely approaches 0 or 1, such that RCP cannot be inferred for completely methylated or unmethylated genomic regions. Nonetheless, RCP can usually be inferred with high confidence using data from only a few hundred dyads. Our new approach therefore requires far fewer sequences to estimate the important parameter of concordance preference than do methods that focus on inferring rates for specific enzyme activities [22, 27].

Here, we apply RCP to investigate further the conclusion of Shipony *et al.* [16] that methylation in cultured stem cells is dominated by non-conservative processes, with little or no preference for concordance. Using double-stranded methylation patterns collected by our group, by Arand *et al.* [14, 17], and by Zhao *et al.* [15], we assess methylation concordance in cultured human and murine stem cells, as well as in murine

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Locus	*G6PD*	*FMR1*				*LEP*			*Lep*
Sample	Lym.	Lym, FX patient	Lym, Normal	FX	IMR90	Leu, M	Leu, F	Adipose	Som, mur.
RCP	13.2	45.2	57.2	63.9	27.2	34.2	85.7	33.8	25.9

Locus	*Afp*	*Igf2*	*Snrpn*	*Tex13*
Sample	Mouse embryonic fibroblasts			
RCP	43.0	4×10^7	65.3	18.2

Figure 2. Inferring RCP for single-copy loci in differentiated human and murine cells. Methylation in human and murine differentiated cells was consistently inferred to have strong contributions from conservative processes for data sets that span a wide range of methylation frequencies. For each locus, we inferred the point estimate of the m, U pair and the two-dimensional 95% confidence region, determined by uncertainty in each variable. The intensity of coloration in a given part of a confidence region reflects the confidence level (α) at that point. The m, U point estimates for most data sets are indicated with white asterisks, and the corresponding RCP values, corrected for bias, are given in the associated table. A larger, colored asterisk is used when the confidence interval of a data set is too small to be readily visible. **(a)** Data from three human loci — *G6PD*, *FMR1*, and *LEP* — and one murine locus — *Lep* — collected using hairpin-bisulfite PCR and dideoxy sequencing, are taken from published [22, 27–29] and previously unpublished (Table S3) data. **(b)** Data from four single-copy loci from murine embryonic fibroblasts — *Afp*, *Igf2*, *Snrpn*, *Tex13* — collected by Arand *et al.* [14] using hairpin-bisulfite PCR and pyrosequencing; relevant values are given in Table S4. Information on the regions of each locus for relevant methylation data are given in Tables S3 and S4. All data were corrected for inappropriate and failed conversion of methylcytosine using methods described in Supporting Information (S4 Text). Conversion-error rate estimates and 95% confidence intervals on RCP are given in Tables S3 and S4.

cells undergoing early developmental transitions that give rise to totipotent embryonic cells.

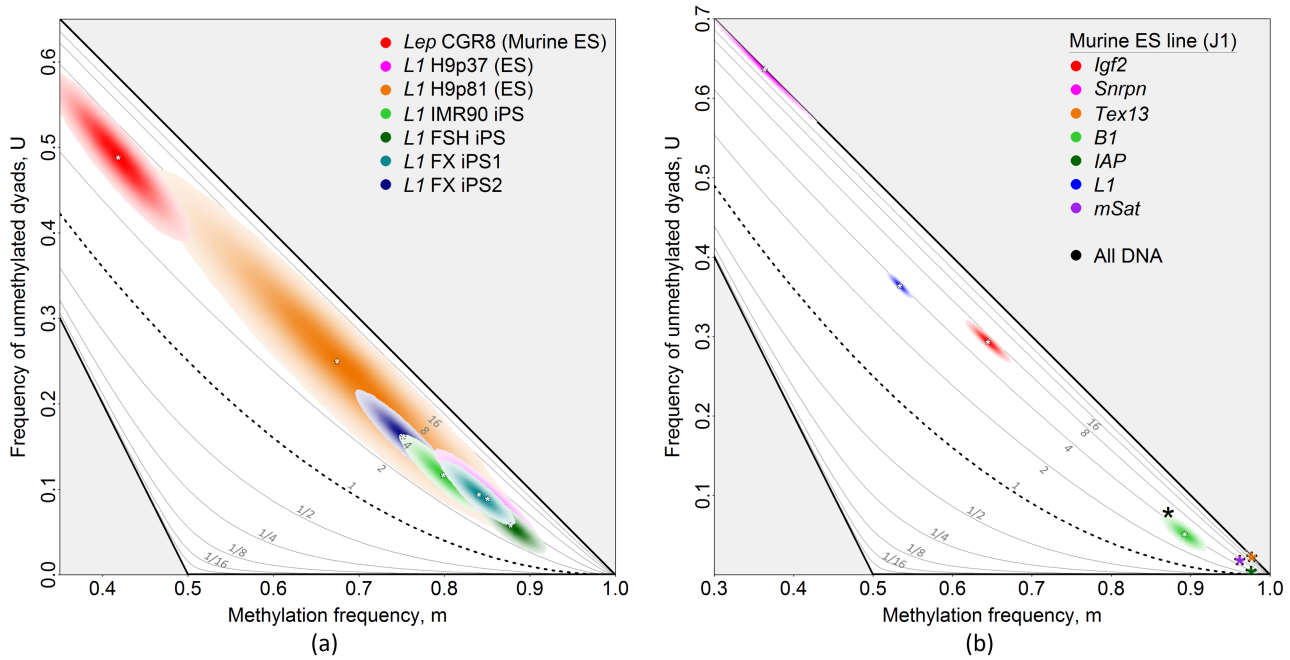
Inferring RCP for Single-Copy Loci in Differentiated Cells

We first examined methylation concordance at single-copy loci in differentiated cells, including those from human adipose, leukocytes, and cultured fibroblasts, and from murine somatic tissues. Our previous work with single-copy loci in differentiated cells has revealed a substantial role for maintenance methylation, a conservative process, with a comparatively minor role for non-conservative *de novo* processes [22, 27]. We therefore anticipated that RCP analysis would indicate substantial preference for

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

concordant methylation states. RCP point estimates for loci in human differentiated cells indicated a 13- to 89-fold preference for concordant methylation (Fig 2a; Table S1; see Table S3 for 95% CI on these point estimates), reflecting, as anticipated, a substantial role for conservative processes. It is noteworthy that the RCP values inferred for these loci have close correspondence with Hemi-Preference Ratio, a metric we have used previously (S2 Text).

A substantial role for conservative methylation processes in murine differentiated cells was also revealed by RCP analysis. Data sets from Stöger [28] for adipose tissue, and from Arand *et al.* [14] for embryonic fibroblasts (MEFs) each gave an RCP point estimate greater than 18.2 (Fig 2a and b). These analyses of single-copy loci in human and murine differentiated cells revealed, in addition, that RCP values indicative of a strong role for conservative processes can be inferred for data sets with disparate methylation frequencies, ranging from sparse to dense (Fig 2).



Locus	*Lep*	*L1*					
Cell Line	CGR8 (Murine ES)	H9p37 (ES)	H9p81 (ES)	IMR90 (iPS)	FSH (iPS)	FX (iPS1)	FX (iPS2)
RCP	4.25	4.41	5.20	3.41	3.41	4.14	3.83

Locus	*Igf2*	*Snrpn*	*Tex13*	*B1*	*IAP*	*L1*	*mSat*	All DNA
RCP	6.72	3×10^8	4×10^8	3.70	3.69	3.88	6.88	5.22

Figure 3. Inferring RCP in human and murine stem cells. Methylation patterns in undifferentiated, cultured human and murine stem cells were consistently inferred to have substantial contributions from conservative processes, with concordance much greater than the random expectation. **(a)** Data from human *L1* and *Lep*, collected using hairpin-bisulfite PCR and dideoxy sequencing, are taken from published [28] and previously unpublished work (Table S3). **(b)** Data from three single-copy loci and four multi-copy loci from murine ES line J1 cells, *Igf2*, *Snrpn*, *Tex13*, *B1*, *IAP*, *L1*, and *mSat*, were collected by Arand *et al.* [14] using hairpin-bisulfite PCR and pyrosequencing (Data in Table S4). "All DNA" data were collected by Zhao *et al.* [15] using hairpin-bisulfite PCR and pyrosequencing, and assess methylation at approximately 17.3% of CpG/CpG dyads in the murine genome (Table S5). All data were corrected for inappropriate and failed conversion of methylcytosine using methods described in Supporting Information (S4 Text). Conversion-error rate estimates and 95% confidence intervals on RCP are given in Tables S3, S4, and S5.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Inferring RCP for Loci and Genome-Wide CpGs in Cultured Stem Cells

We next analyzed concordance preference in undifferentiated, cultured stem cells to assess with RCP the conclusion of Shipony *et al.* [16] that methylation in such cells occurs primarily through non-conservative, random methylation. All six of the human stem and iPS cell lines that we examined, when grown under non-differentiating conditions, gave RCP point estimates within the narrow range of 3.41 to 5.20 (CI given in Table S3) for the 5'-UTR of the multi-copy locus *L1*, showing little evidence of heterogeneity in RCP ($p > 0.29$, pairwise two-tailed permutation tests (PTs); Fig 3a; Table S3; further discussion in S5 Text). The single-copy *Lep* locus in a cultured murine ES line [30] yielded an RCP within this range (Fig 3a), and not significantly different from that observed for *L1* transposable elements in each of the cultured human stem cell lines ($p > 0.47$, pairwise two-tailed PTs). The inferred RCP values in this narrow range were all significantly higher than the RCP of 1 expected for the random methylation proposed by Shipony *et al.* ($p < 10^{-4}$, one-tailed bootstrap test).

To extend our analysis of methylation in stem cells, we evaluated published data from additional loci [14], CpGs across the genome, and other classes of genomic elements [15] from cultured murine ES cells. Three of the multi-copy loci — *B1*, *IAP*, and *L1* — had RCP values within the narrow range described above (Fig 3b), with no evidence of heterogeneity among the three loci ($p > 0.55$, pairwise two-tailed PTs; inferred from data of Arand *et al.* [14]). One multi-copy locus — *mSat* — yielded a significantly higher RCP value ($p < 0.017$, pairwise two-tailed PTs). Single-copy loci — *Igf2*, *Snrpn*, and *Tex13* — also had significantly higher RCP values ($p < 0.019$, pairwise two-tailed PTs).

Near-genome-wide double-stranded methylation data from Zhao *et al.* [15], gave an inferred RCP of 5.22 for "all CpGs" in DNA from undifferentiated, cultured murine ES cells. For other classes of genomic elements in this genome-scale study [15], we inferred RCP values of 4.31 or greater (Table S5). In all cases, RCP values were significantly greater than the value of 1 expected under random methylation processes (RCP ≥ 3.69; $p < 10^{-6}$, maximum-likelihood comparison test (MLCT); Fig 3b). Thus, RCP values greater than 1 are a consistent feature of cultured embryonic stem cells, and hold across a broad set of genomic locations and element categories.

We considered the possibility that spontaneous differentiation had produced subpopulations of cultured stem cells that might account for the inference of RCP values substantially greater than 1 at all loci and genomic elements examined. Our calculations revealed that a possible subpopulation of differentiated cells operating at much higher RCP than that of undifferentiated cells would need to comprise more than 50% of the population to account for our finding (S9 Text). Morphological inspection of the cultured human stem cells under non-differentiating conditions did not suggest the presence of a substantial subpopulation of differentiated cells in any of these lines. We conclude that overall methylation in cultured mouse and human ES and iPS cells is established with appreciable preference for concordance.

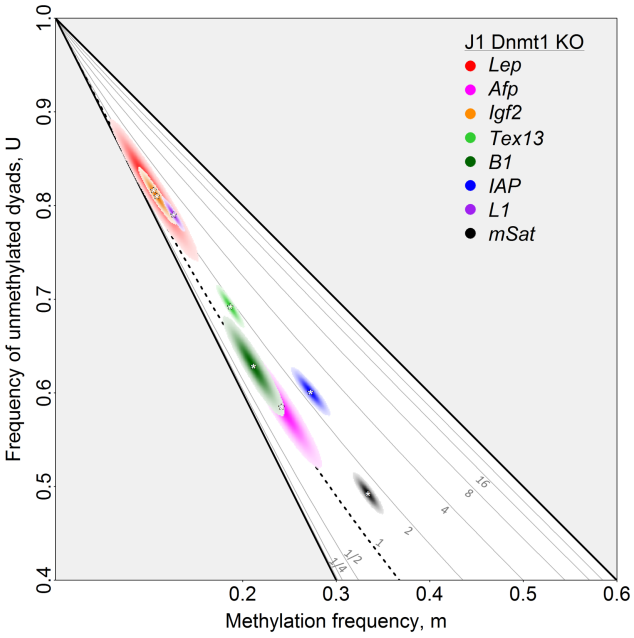
Inferring RCP in *Dnmt1*-Knockout ES Cells

Our finding of substantial preference for methylation concordance in these data from cultured, undifferentiated stem cells contrasts with the inference of Shipony *et al.* [16]

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that DNA methylation in such cells is dominated by non-conservative, random processes. This disparity led us to ask whether our approach here for data acquisition and analysis is indeed capable of identifying data sets established under exclusively random processes, which are expected to yield RCP values of 1 (see "Defining the Ratio ...", above). To assess this capacity, we considered data collected from the murine *Dnmt* and Np95 knockout (KO) lines, and compared RCP values at various loci to those for normal murine ES and differentiated cells [14, 28, 30].

The Dnmt1KO cell line is of particular interest in this regard because Dnmt1, known as the maintenance methyltransferase, has the strongest known preference for establishing concordant methylation [11, 22, 31]. For murine ES cells homozygous for a knockout at *Dnmt1*, or for a knockout mutation at *Np95*, which encodes a protein critical for Dnmt1 maintenance activity [32], we inferred a range of RCP point estimates of 1.02 to 2.35 for repeated elements and single-copy loci from Arand *et al.* [14] (RCP values inferred for Dnmt1KO data shown in Fig 4; CI for both mutant lines given in Table S4). Similarly, data from Al-Azzawi *et al.* yield a point estimate of 1.90 for the *Lep* locus in murine cells homozygous for a Dnmt1KO [30] (Fig 4; CI given in Table S3). DnmtKO and Np95KO cell lines had RCP values less than those inferred for wildtype murine stem cells at all loci examined ($p < 0.049$, two-tailed PTs; Table S4). While for ten of thirteen data sets examined, RCP values were inferred to be greater than 1, in three notable instances, RCP values were not significantly greater than 1, indicating methylation configurations consistent with expectations under wholly random placement of methyl groups ($p = 0.17$ and $p = 0.10$ for *B1* and *Afp* in Dnmt1KO cells



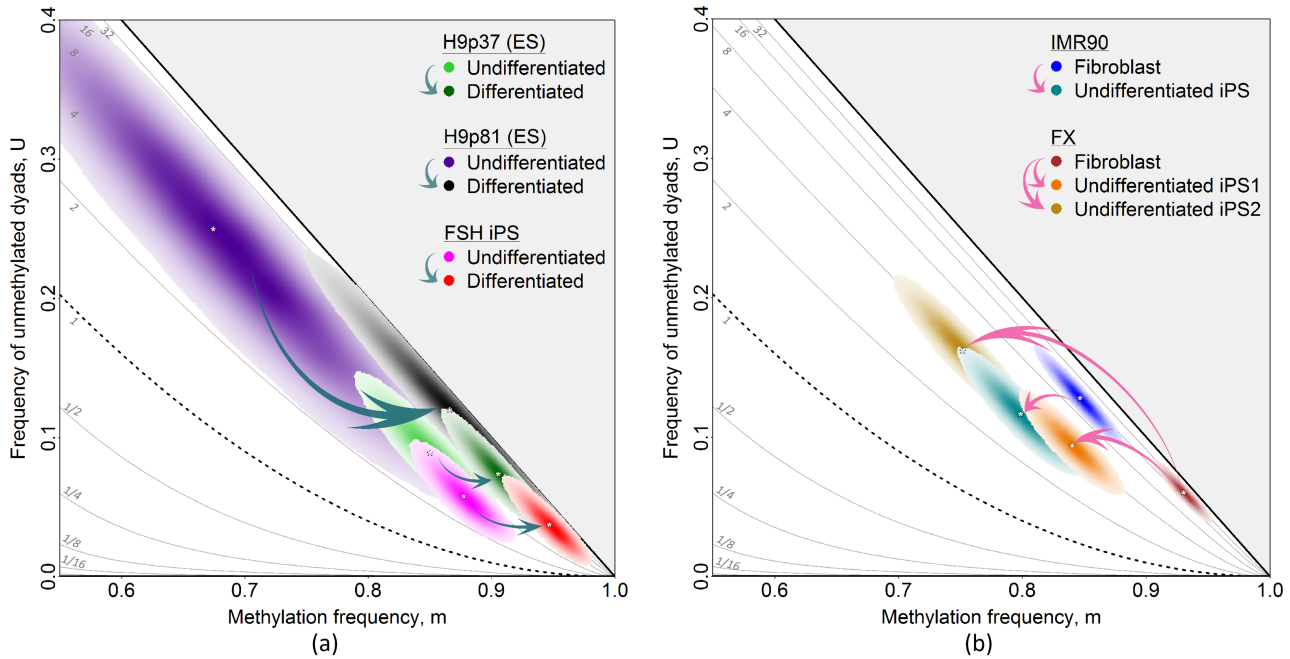
Locus	Cell Line	RCP
Lep	J1 Dnmt1 KO	1.90
Afp		1.16
Igf2		1.83
Tex13		1.77
B1		1.14
IAP		2.35
L1		2.22
mSat		1.62

Figure 4. Inferring RCP in *Dnmt1*-knockout ES cells. In the absence of *Dnmt1*, RCPs inferred for cultured murine ES cells were close to 1, the random expectation, at some loci. Data from *Lep*, collected using hairpin-bisulfite PCR and dideoxy sequencing, are from Al-Alzzawi *et al.* [30]. Data from seven additional loci are from Arand *et al.* [14], who used hairpin-bisulfite PCR and pyrosequencing. Our analysis of these published data revealed two of the eight loci analyzed — *Afp* and *B1* — to have very low RCP values whose 95% confidence intervals include the possibility that methylation was established entirely by random processes. Data were corrected prior to RCP analysis for inappropriate and failed conversion of methylcytosine using methods described in Supporting Information (S4 Text). Conversion-error rate estimates and 95% confidence intervals on RCP are given in Tables S3 and S4.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

respectively, and $p = 0.38$ for *B1* in Np95KO cells, one-tailed bootstrap tests; Table S4). These findings confirm the ability of the RCP framework to detect methylation established with little or no preference for concordance or discordance, and strengthen our conclusion that cultured stem cells do have substantial contributions from conservative DNA methylation processes (Fig 3).

Our inference of some RCP values slightly but significantly greater than 1 in cells deprived of Dnmt1 or its activity hints at some residual maintenance activity contributed by the Dnmt3 methyltransferases. We inferred a complex role for Dnmt3s in our analysis of data from Dnmt3KO lines (S10 Text), cconsistent with earlier suggestions [19, 22, 33] that the contributions of the Dnmt3s can include low levels of maintenance activity.



	U→D			fibro → iPS		
Cell line	H9p37	H9p81	FSH iPS	IMR90	FX	
Undifferentiated	4.41	5.20	3.41	3.41	4.14	3.83
Differentiated	12.3	23.2	12.3	12.8	25.3	

Figure 5. Shifts in RCP of *L1* elements upon differentiation and dedifferentiation in cultured human ES and iPS cells. We compare RCP values for *L1* elements in cultured stem cells grown under non-differentiating and differentiating conditions, and, for iPS cells, their progenitor fibroblast lines. Blue arrows indicate differentiation and pink arrows indicate dedifferentiation. **(a)** For each of three human ES and iPS lines, undifferentiated cells had only moderate preference for concordance. Upon differentiation, RCP values for the cell lines shifted in parallel toward stronger preference for concordance. **(b)** For two differentiated human fibroblast lines, dedifferentiation to iPS cells resulted in a shift to reduced RCP values, indicating lower preference for concordance. Note that one fibroblast line (FX) was used to establish two different iPS lines, which differed in their methylation frequency, m, at *L1* elements, but had similar RCP values. Some data points were previously shown in Fig 2 and Fig 3, and are included here to illustrate more directly the relationships of the differentiated and undifferentiated cultured cells. Conversion-error rate estimates and 95% confidence intervals on RCP are given in Tables S3.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Shifts in RCP Upon Differentiation and Dedifferentiation of Cultured Cells

Our initial examination of RCP values in differentiated cells as compared to cultured stem cells suggested that RCP is altered through the differentiation process (Fig 2 and Fig 3). Would significant RCP increases be observed for individual cell lines upon transition in culture from one differentiation state to another? We first asked whether RCP values change when undifferentiated human ES and iPS cells are grown under differentiating conditions (see Materials and Methods). We inferred RCP at the promoter of *L1* elements of cultured human iPS and ES cells, with two different passages for the latter. Upon differentiation, RCP values for all three of these cell lines increased significantly ($p = 0.0004$, H9p37; $p = 0.025$, H9p81; $p = 0.0006$, FSH iPS; two-tailed PTs), and approached the lower boundary of the confidence region inferred for single-copy loci in differentiated somatic cells (Fig 2 and Fig 5a). Using near-genome-wide data for cultured murine cells [15], we inferred significant RCP increases upon cell differentiation for most genomic elements ($p < 10^{-20}$, MLCTs), with the exception of low-complexity and satellite DNAs (Table S5). These RCP increases were greatest at promoters, CG islands, and CG island shores, and are more muted at other regions. We conclude that the onset of differentiation in cultured human and murine cells is usually associated with a shift towards a greater role for conservative processes.

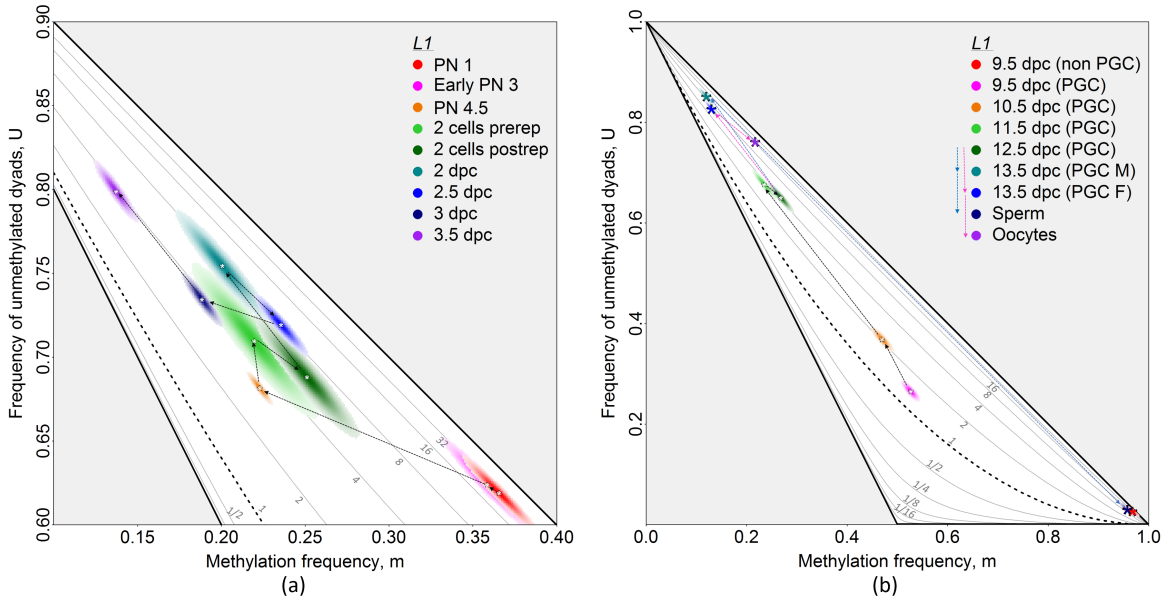
We then asked if the dedifferentiation that occurs upon production of an iPS line from a differentiated cell would have opposite effects on concordance preference. We compared methylation at *L1* elements in three iPS lines to methylation in the two cultured human fibroblast lines from which they were derived. As predicted, RCP values for all three iPS lines were much reduced compared with values observed for the parent fibroblast lines ($p < 10^{-5}$, two-tailed PTs; Fig 5b). Dedifferentiation in tissue culture is thus associated with a shift in DNA methylation toward a greater role for non-conservative processes. It will be useful to investigate whether changes in methylation systems as measured by RCP drive or merely reflect the cellular differentiation process.

Changes in RCP During Early Embryonic and Primordial Germ-Cell Development

Our inference of ~4-fold-or-greater preference for concordance in many different cultured ES and iPS cell lines (Fig 3a; Table S3) indicates that it is possible for cultured stem cells to have appreciable levels of conservative methylation. We next asked whether uncultured stem cells taken directly from an embryo also have strong contributions from conservative methylation processes. In so doing, we can also ask whether totipotent stem cells are similar to pluripotent stem cells in their levels of conservative methylation.

To investigate these questions, we applied RCP to double-stranded methylation patterns collected by Arand *et al.* [17] for three multi-copy loci in mouse embryos, *L1*, *mSat*, and *IAP*. Our analyses above indicate that, in general, multi-copy loci have RCP values in cultured stem cells that are lower than those for single-copy loci, and are similar to or lower than many other genomic elements (Figs 2, 3, 4, and Tables S3, S4, S5). Analysis of the three multi-copy loci is therefore expected to provide a conservative estimate of the concordance preference of broader genomic regions.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Stage	PN 1	Early PN 3	PN 4-5	2 cells prerep	2 cells postrep	2 dpc	2.5 dpc	3 dpc	3.5 dpc
RCP	30.6	26.6	3.09	4.61	5.95	7.64	8.20	3.71	3.80

	Non-PGC	PGC						Germ cells	
Stage (dpc)	9.5	9.5	10.5	11.5	12.5	13.5 (M)	13.5 (F)	Sperm	Oocytes
RCP	26.0	1.40	2.08	3.57	4.27	9.77	6.17	13.4	17.6

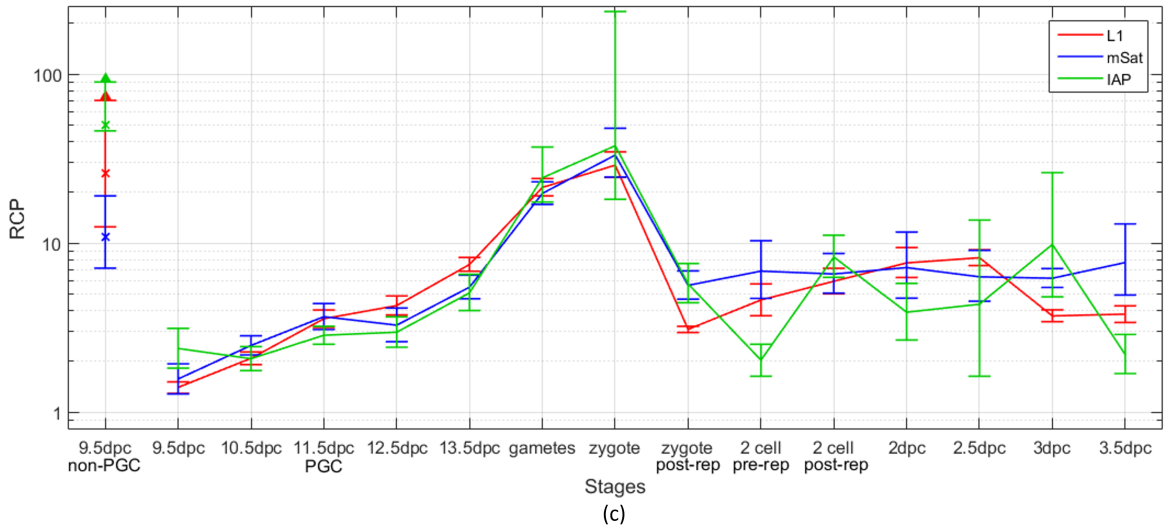


Figure 6. Shifts in RCP during embryonic and germ-cell development. Major transitions in RCP values occur during early embryonic and primordial germ cell development. RCP values were calculated from data on DNA methylation at *L1* elements in Arand *et al.* [17]. For both early embryonic stages and primordial germ cell development, progressions over time in m, U, and subsequently RCP values, are indicated with arrows. In cases where data were available for multiple replicates, the dyad frequencies were pooled. Corresponding estimates of RCP, corrected for bias, are given in the associated tables and in (c). **(a)** Transitions from early pronuclear stages (1 and 3) to late stages (4-5) were accompanied by a sharp decrease in RCP, to a level similar to that observed in cultured stem cells (Fig 3). Further embryonic development was accompanied by minor increases and subsequent decreases in methylation and RCP values. **(b)** Primordial germ cells (PGCs) at the earliest stage for which data are available, 9.5 days post conception (dpc), had RCP values at *L1* elements even lower than those inferred for Dnmt1-KO cells. RCP values increased during PGC maturation to stage 13.5 dpc even as methylation frequencies decreased by more than 50%. For differentiated gametes, RCP values were similar to those found in other differentiated cells (see Fig 2), while methylation frequencies differed markedly between eggs and sperm. **(c)** Tracking RCP at multi-copy loci through early embryonic development. We inferred RCP for the various embryonic stages for which methylation data were reported by Arand *et al.* [17]. Our figure parallels their Fig 6b, which used the percentage of hemimethylated CpG dyads relative to all methylated CpG dyads, rather than the RCPs inferred here. Point estimates and confidence intervals (95%) for all but one data set were estimated by bootstrapping and applying the BCa correction (S7 Text). Double-stranded sequences were unavailable for 3 dpc at *mSat*, and thus a likelihood-based method was used for the estimation of its confidence intervals (S8 Text). Conversion-error rate estimates and 95% confidence intervals on RCP are given in Tables S4.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Our analyses for these three multi-copy loci revealed that totipotent embryonic cells, taken from the two-cell to the morula stages (through 3 days post conception (dpc)), also exhibit substantial preference for concordance, and hence, conservative methylation. RCP values ranged from 3.71 to 9.83 over this interval (Fig 6a; Table S4), with the exception of a transiently lower RCP value of 2.03 for the *IAP* locus in the prereplicative two-cell zygote (Table S4). Data for uncultured pluripotent stem cells from mouse embryos (gastrula, 3.5 dpc), although considerably more limited than for totipotent stem cells, indicate similarly strong preferences for concordant methylation at two of the three multi-copy loci examined, with *IAP* again being the exception. Thus, relative to random expectations, uncultured pluripotent and totipotent cells from mouse embryos — with only transient exceptions at one locus — do have substantial preferences for concordance. This echoes our findings for cultured pluripotent murine and human ES and iPS cells. We conclude that the level of preference for concordant methylation exists in the embryo and arose neither in the transition from totipotency to pluripotency nor in the establishment and culturing of stem cells (Fig 3).

RCP values for specific loci and for most genomic regions are consistently lower for stem cells than for differentiated cells (Fig 2 and Fig 5). Do the lower RCP values for stem cells originate in gametes, or do they arise post-fertilization? To address this question, we inferred RCP values for sperm and oocytes, using data from Arand *et al.* [17]. At the three multi-copy loci for which data are available, RCP values for gametes are within the range found for loci in other differentiated cells, with average point estimates ranging from 13.4 to 45.0 (Fig 6b; Table S4). This high preference for concordance in murine sperm and oocytes implies that the lower RCP values characteristic of zygotes must arise during post-fertilization events.

To identify the timing of the transition to the lower RCP values observed in stem cells, we next assessed data from post-fertilization nuclei and cells [17] . Pronuclear stages 1 and 3 revealed high RCP values, similar to those observed in gametes. In pronuclear stages 4-5, however, there was an abrupt transition to lower RCP values in the range observed for totipotent stem cells (Fig 6).

Is this transition dependent on a DNA replication event that occurs between pronuclear stages 3 and 4-5? To address this question, we assessed data from Arand *et al.* [17], in which aphidicolin was used to block DNA replication in the fertilized egg. Our RCP analysis reveals that methylation patterns at *L1* and *mSat* in these treated nuclei, while having somewhat lower RCP values relative to pronuclei at earlier stages, had not undergone the major reduction in RCP that we inferred for unmanipulated pronuclei at stages 4-5 ($p < 10^{-5}$, two-tailed PTs; Table S4). Thus, the shift to lower RCP in stem cells following fertilization appears to require either DNA replication or some later event that is itself replication-dependent. Our conclusion that pronuclear replication is pivotal to marked changes in RCP is consistent with the inference of Arand *et al.*, using a different metric [17], that this replication is critical to changes in methylation dynamics.

The change from high RCP values in gametes and early pronuclei to lower values in stem cells was markedly abrupt. Are other major transitions in RCP similarly abrupt, or do some occur gradually, perhaps over many cell divisions? Murine primordial germ cells (PGCs), in their progression to mature gametes, offer an opportunity to approach this question [34,35]. Using data from Arand *et al.* [17], we inferred that RCP values for PGCs increased by factors ranging from 2.5 to 5 during their progression from 9.5 dpc to 13.5 dpc. This increase was not sudden but gradual over the four-day period, spanning an interval of substantial cell proliferation [34] (Fig 6b). The murine embryo data thus provide examples of both abrupt and gradual transitions in RCP through

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

development (Fig 6c).

The RCP values for PGCs at 9.5 dpc, the earliest stage for which data were reported by Arand *et al.* [17], were strikingly low. For all three loci studied, RCPs — 1.40 for *L1*, 1.57 for *mSat*, and 2.38 for *IAP* — were either not significantly different from or lower than those inferred for the same loci in the Dnmt1-KO line (*L1*: $p < 10^{-5}$; *mSat*: $p = 0.78$; *IAP*: $p = 0.93$; two-tailed PTs; Fig 4). These were among the lowest RCP values we inferred for the developing mouse embryos, and perhaps reflect the epigenetic flexibility required for the production of gametes.

Over the first few rounds of PGC division, developmental reprogramming and commitment are occurring [35], and during this time lineage-specific gene expression patterns are arising. The gradual increase in RCP observed across division of these PGCs may mirror the increases in RCP values that occur when cultured ES cells are subjected to differentiating conditions (Fig 5).

There is, however, a critical difference between the trajectories for proliferating PGCs and differentiating ES and iPS cells in culture. When cultured cells were differentiated, their methylation frequencies increased. By contrast, methylation frequencies for PGCs declined across rounds of proliferation [36]. Seisenberger *et al.* [7], Arand *et al.* [17], and von Meyenn *et al.* [18] concluded that this reduction in methylation frequency is driven by partial impairment of maintenance methylation.

Our RCP framework permits a closer look at the likely extent of this proposed maintenance impairment. If maintenance methylation were completely impaired and no other methylation processes were active, the resulting passive demethylation would be fully dispersive: methyl groups would remain only in hemimethylated dyads. A fully dispersive process such as this should halve the methylation frequency at each DNA replication event and yield an RCP value of 0. Such extremely low values of RCP were inferred only for cells treated with S-Adenosylmethionine-ase (Table S4). As this treatment either impairs or eliminates a cell's ability to methylate DNA, it reveals RCP trajectories that would be observed with complete or nearly complete suspension of all methylation processes. Our finding of RCP values greater than 0, as well as 1, suggests not only that conservative processes remain in maturing PGCs, but also that their contributions are sufficient to outweigh any dispersive effects of passive demethylation.

Concluding Remarks

The Ratio of Concordance Preference, RCP, quantifies the degree to which exact information is conserved when binary information is transferred in a given system. With an explicit null model for the transfer of information, RCP is a scale for assessing observed preference for concordance as compared to a random expectation, such as the random placement of methyl groups. It thus permits quantitative comparison of information-transfer systems that have disparate binary-state frequencies. Here, we have applied RCP to double-stranded CpG DNA methylation data from cultured and embryonic cells to track and to understand the shifts in their strategies for information transfer, with particular focus on changes across differentiation states and developmental time. The mechanism-free nature of RCP allows the comparison of this DNA methylation property not only across different cell types and developmental stages, but also across organisms whose methylation machineries may differ significantly.

Application of RCP to double-stranded DNA methylation patterns enabled us to

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

answer three questions raised above: (i) in cultured pluripotent and in embryonic totipotent stem cells, preference for concordance is reduced but not eliminated compared with differentiated cells; (ii) cellular differentiation for some classes of genomic elements such as promoters is characterized by increasing preference for concordance, whereas dedifferentiation reflects a decline in preference for concordance; and (iii) concordance preference decreases markedly following fertilization and the first round of DNA replication, but remains substantial through totipotency and early stages of pluripotency; at three multi-copy loci, primordial germ cells initially have nearly no preference for concordance, perhaps reflecting the high level of epigenetic flexibility *en route* to production of functional gametes.

RCP analysis of cultured pluripotent and embryonic totipotent cells provided an opportunity to reassess the balance between conservative and non-conservative processes that contributed to the "noisy" methylation patterns described by Shipony *et al.* [16]. We found evidence of substantial contributions from conservative processes in stem cells — both cultured pluripotent and embryonic totipotent — using a variety of single- and multi-copy loci, and near-genome-wide data for all-CpG methylation. Our results are thus contrary to the conclusion that "dynamic" processes dominate DNA methylation in stem cells [16].

The disparity between the proposal of dynamic methylation by Shipony *et al.* [16] and our inference of appreciable conservative methylation in cultured stem cells may have arisen from methodological differences in data collection. Inappropriate conversion of methylcytosines can occur during prolonged bisulfite treatment [37], producing artifactual variation among observed patterns [38] (S3 Text). Under methods that quantify diversity among single-stranded patterns but do not account for inappropriate conversion, such errors will lead to overestimation of the role of random as compared to conservative processes.

Another possible explanation for this disparity is that the separation and culturing of individual ES and iPS cells, a procedural step used by Shipony *et al.* [16], leads to unusually diverse patterns of DNA methylation among descendant cells. Thus, it is possible that low preference for concordance can occur in stem cells that are isolated from other such cells, including, perhaps, PGCs. This possibility, if correct, would offer new opportunities in stem-cell biology.

Regardless of the possible impacts of stem cell isolation and of conversion error on discrepancies in data, the proposal of Shipony *et al.* [16] that developmental reprogramming occurs with a high level of epigenetic flexibility is interesting and merits consideration. Our observation that very early primordial germ cells have RCP values close to the random expectation could reflect such flexibility *en route* to production of functional gametes. Future experiments, however, will be necessary to determine whether our inference of an RCP value close to 1 for early PGCs is a consequence of almost entirely random processes, or whether it instead results from a balance of conservative and dispersive methylation, consistent with the interpretations of Arand *et al.* [17] and von Meyenn *et al.* [18].

We have observed that loci in a given cell type can have disparate RCP values, with the trajectories of RCP changing through time. Significant differences in RCP were observed among loci in individual lines of cultured ES and iPS, and embryonic cells, with particularly pronounced differences between some single- and multi-copy loci (Fig 3b). Furthermore, while the three multi-copy loci we analyzed in data from developing embryos yielded relatively consistent inferences of RCP for each stage, one

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

locus was observed to show larger RCP shifts during and near totipotency (Fig 6c). Locus-specific features such as imprinting and chromosome location likely contribute to some of this variation. Differences in developmental trajectories of flexibility and stability of DNA methylation may correspond to variation among loci in the timing of their epigenetic sensitivity to environmental conditions. By contrast, the grouping of genome-wide data [15] into functional classes (Table S3) may tend to mute the observed RCP changes during development, as members of each class are likely to differ in their involvement in the processes of differentiation.

DNA sequencing technologies are beginning to yield longer and more informative double-stranded DNA methylation patterns without the need for bisulfite conversion and PCR amplification [39, 40]. Such methods can reduce data corruption that arises through conversion and amplification errors, and may also enable distinction between methyl- and hydroxy-methyl cytosine. Longer reads may make it possible to assess and compare concordance preferences for individual cells. Little is known of the variation in epigenetic processes among cells from seemingly homogeneous populations. Many human cancers and syndromes are characterized by a combination of genetic and epigenetic changes [2, 29, 41]. Assessing the mosaicism of single-cell RCP values in such cases may lead to deeper understanding of the origin and nature of epigenetic heterogeneity in human disease.

Materials and Methods

Mathematical foundations of the RCP framework

Overview of the mathematical foundation is given in Results and Discussion, and developed further in S1 Text.

Human Cells and Culture Conditions

The six human ES and iPS cell lines for which we collected methylation patterns were derived from either embryos or fibroblasts described as normal [42] or from fibroblasts of individuals with disorders not known to affect the basic biochemistry of DNA methylation.

Many of the sets of human DNA methylation patterns analyzed here were presented in previous publications, which include information on University of Washington Human Subjects approval for collection and use. These data include *G6PD* and *FMR1* from leukocytes of normal individuals [22, 27]; *FMR1* from males with fragile X syndrome [29]; *LEP* from male leukocytes and from female lymphocytes and adipose tissue [22, 28].

Human methylation patterns presented here for the first time were collected from: (i) FX iPS cell lines 1 and 2, which were developed at the University of Washington ISCRM facility from fibroblasts (line GM07730, Coriell Cell Repositories, Camden, NJ) of a male with a fragile X "full mutation", using published methods [43] ; (ii) iPS cell line IMR90, which was developed at the University of Washington ISCRM facility from the IMR90 somatic line established from fibroblasts (obtained from ATCC) of a normal female, using published methods [43]; (iii) FSH iPS cell line, which was developed at the

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

University of Washington ISCRM facility from fibroblasts of an individual with Facioscapulohumeral dystrophy (FSHD), using published methods [44]; and (iv) H9 human ES cells from NIH Embryonic Stem Cell Registry (WA09, H9 number 0062).

Cells were cultured in Dulbecco's modified Eagle's medium/Ham's F-12 medium containing GlutaMax supplemented with 20 percent serum replacer (SR), 1 mM sodium pyruvate, 0.1 mM nonessential amino acids, 50 U/ml penicillin, 50 μg/ml streptomycin and 10 ng/ml basic fibroblast growth factor (all from Invitrogen), and 0.1 mM β-mercaptoethanol (Sigma-Aldrich). hESCs were grown on γ-irradiated primary mouse embryonic fibroblasts (MEFs) and passaged using dispase (1.2 U/ml; Invitrogen). They were passaged onto Matrigel (Corning) without feeders in mTeSR1 (Stem Cell Technologies) for the final passages prior to analysis. Cells were differentiated by passage onto Matrigel in Dulbecco's modified Eagle's Medium supplemented with 20 percent fetal bovine serum and pen/strep. Images of our cultured stem cells grown under differentiating conditions confirmed their pluripotency.

Murine Cells and Culture Conditions

Methylation patterns from murine ES cells, and the origin and culturing of these cells, have previously been described [14, 15, 17, 30].

Collection of Double-Stranded DNA methylation patterns using hairpin-bisulfite PCR

The DNA methylation patterns collected in our lab and analyzed here, both those published previously and those presented here for the first time, were collected using the hairpin-bisulfite PCR approach [13], with barcodes and batchstamps to authenticate each sequence [45]. Details for collection of each data set are given in Table S2.

The data presented by Arand *et al.*] [14, 17], and Zhao *et al.*, [15] and analyzed here, were collected in the absence of molecular batch-stamps and barcodes, raising the possibilty that the reliably of those data sets is undermined by PCR clonality. However, both groups used alternate strategies that revealed that PCR clonality was not rampant. Zhao *et al.* found that essential features of data sets did not differ appreciably between the "real" data set collected conventionally, using PCR, and a test, PCR-free data set that excluded opportunities for clonality by including only one read from each locus, providing no evidence of impacts from PCR clonality (Hehuang Xie, personal communication). In turn, Arand *et al.* used molecular codes for several of their data sets, and, for the few data sets collected in the absence of such code, observed appreciable heterogeneity among patterns, also hinting that data were not appreciably impacted by clonality (Julia Arand, personal communication).

References

1. Feinberg AP, Vogelstein B, et al. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983;301(5895):89–92.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2. Baylin SB, Jones PA. A decade of exploring the cancer epigenome—biological and translational implications. Nature Reviews Cancer. 2011;11(10):726–734.

3. Chen Y, Breeze CE, Zhen S, Beck S, Teschendorff AE. Tissue-independent and tissue-specific patterns of DNA methylation alteration in cancer. Epigenetics & chromatin. 2016;9(1):1.

4. Ohi Y, Qin H, Hong C, Blouin L, Polo JM, Guo T, et al. Incomplete DNA methylation underlies a transcriptional memory of somatic cells in human iPS cells. Nature cell biology. 2011;13(5):541–549.

5. Monk M, Boubelik M, Lehnert S. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development. 1987;99(3):371–382.

6. Santos F, Hendrich B, Reik W, Dean W. Dynamic reprogramming of DNA methylation in the early mouse embryo. Developmental biology. 2002;241(1):172–182.

7. Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, et al. The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. Molecular cell. 2012;48(6):849–862.

8. Guo JU, Su Y, Zhong C, Ming Gl, Song H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell. 2011;145(3):423–434.

9. Bird AP, Southern EM. Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from *Xenopus laevis*. Journal of Molecular Biology. 1978;118(1):27–47.

10. Bestor T, Laudano A, Mattaliano R, Ingram V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells: the carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. Journal of Molecular Biology. 1988;203(4):971–983.

11. Smith SS, Kaplan BE, Sowers LC, Newman EM. Mechanism of human methyl-directed DNA methyltransferase and the fidelity of cytosine methylation. Proceedings of the National Academy of Sciences, USA. 1992;89(10):4744–4748.

12. Kumar S, Cheng X, Klimasauskas S, Mi S, Posfai J, Roberts RJ, et al. The DNA (cytosine-5) methyltransferases. Nucleic Acids Research. 1994;22(1):1.

13. Laird CD, Pleasant ND, Clark AD, Sneeden JL, Hassan KA, Manley NC, et al. Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. Proceedings of the National Academy of Sciences, USA. 2004;101(1):204–209.

14. Arand J, Spieler D, Karius T, Branco MR, Meilinger D, Meissner A, et al. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. PLoS Genetics. 2012;8(6):e1002750.

15. Zhao L, Sun Ma, Li Z, Bai X, Yu M, Wang M, et al. The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. Genome Research. 2014;p. 1296–1307.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

16. Shipony Z, Mukamel Z, Cohen NM, Landan G, Chomsky E, Zeliger SR, et al. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. Nature. 2014;513(7516):115–119.

17. Arand J, Wossidlo M, Lepikhov K, Peat JR, Reik W, Walter J. Selective impairment of methylation maintenance is the major cause of DNA methylation reprogramming in the early embryo. Epigenetics & Chromatin. 2015;8(1):1.

18. von Meyenn F, Iurlaro M, Habibi E, Liu NQ, Salehzadeh-Yazdi A, Santos F, et al. Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells. Molecular Cell. 2016;in press.

19. Liang G, Chan MF, Tomigahara Y, Tsai YC, Gonzales FA, Li E, et al. Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Molecular and Cellular Biology. 2002;22(2):480–491.

20. Jeltsch A, Jurkowska RZ. New concepts in DNA methylation. Trends in Biochemical Sciences. 2014;39(7):310–318.

21. Gowher H, Jeltsch A. Molecular enzymology of the catalytic domains of the Dnmt3a and Dnmt3b DNA methyltransferases. Journal of Biological Chemistry. 2002;277(23):20409–20414.

22. Fu AQ, Genereux DP, Stöger R, Burden AF, Laird CD, Stephens M. Statistical inference of in vivo properties of human DNA methyltransferases from double-stranded methylation patterns. PLoS One. 2012;7(3):e32225.

23. Weinberg W. Über vererbungsgesetze beim menschen. Molecular and General Genetics MGG. 1908;1(1):440–460.

24. Hardy GH. Mendelian proportions in a mixed population. Science. 1908;p. 49–50.

25. Russell DW, Zinder ND. Hemimethylation prevents DNA replication in *E. coli*. Cell. 1987;50(7):1071–1079.

26. Shakibai N, Ishidate K, Reshetnyak E, Gunji S, Kohiyama M, Rothfield L. High-affinity binding of hemimethylated oriC by *Escherichia coli* membranes is mediated by a multiprotein system that includes SeqA and a newly identified factor, SeqB. Proceedings of the National Academy of Sciences, USA. 1998;95(19):11117–11121.

27. Fu AQ, Genereux DP, Stöger R, Laird CD, Stephens M. Statistical inference of transmission fidelity of DNA methylation patterns over somatic cell divisions in mammals. The Annals of Applied Statistics. 2010;4(2):871.

28. Stöger R. In vivo methylation patterns of the *leptin* promoter in human and mouse. Epigenetics. 2006;1(4):155–162.

29. Stöger R, Genereux DP, Hagerman RJ, Hagerman PJ, Tassone F, Laird CD. Testing the FMR1 promoter for mosaicism in DNA methylation among CpG sites, strands, and cells in FMR1-expressing males with fragile X syndrome. PLoS One. 2011;6(8):e23648.

30. Al-Azzawi H. Cytosine methylation and hydroxymethylation at the *Leptin* promoter. Ph.D. dissertation, University of Nottingham; 2013.

31. Jeltsch A. On the enzymatic properties of Dnmt1: specificity, processivity, mechanism of linear diffusion and allosteric regulation of the enzyme. Epigenetics. 2006;1(2):63–66.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

32. Sharif J, Muto M, Takebayashi Si, Suetake I, Iwamatsu A, Endo TA, et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature. 2007;450(7168).

33. Vilkaitis G, Suetake I, Klimašauskas S, Tajima S. Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. Journal of Biological Chemistry. 2005;280(1):64–72.

34. Mintz B, Russell ES. Gene-induced embryological modifications of primordial germ cells in the mouse. Journal of Experimental Zoology. 1957;134(2):207–237.

35. McLaren A. Primordial germ cells in the mouse. Developmental biology. 2003;262(1):1–15.

36. Tang WW, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A unique gene regulatory network resets the human germline epigenome for development. Cell. 2015;161(6):1453–1467.

37. Hayatsu H, Negishi K, Shiraishi M. Accelerated bisulfite-deamination of cytosine in the genomic sequencing procedure for DNA methylation analysis. In: Nucleic Acids Symposium Series. vol. 48. Oxford Univ Press; 2004. p. 261–262.

38. Genereux DP, Johnson WC, Burden AF, Stöger R, Laird CD. Errors in the bisulfite conversion of DNA: modulating inappropriate-and failed-conversion frequencies. Nucleic Acids Research. 2008;36(22):e150.

39. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13(1):341.

40. Laszlo AH, Derrington IM, Brinkerhoff H, Langford KW, Nova IC, Samson JM, et al. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. Proceedings of the National Academy of Sciences, USA. 2013;110(47):18904–18909.

41. Lemmers RJ, Tawil R, Petek LM, Balog J, Block GJ, Santen GW, et al. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nature genetics. 2012;44(12):1370–1374.

42. Thomson JA, Itskovitz-Eldor J, Shapiro SS, Waknitz MA, Swiergiel JJ, Marshall VS, et al. Embryonic stem cell lines derived from human blastocysts. Science. 1998;282(5391):1145–1147.

43. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, et al. Induced pluripotent stem cell lines derived from human somatic cells. Science. 2007;318(5858):1917–1920.

44. Snider L, Geng LN, Lemmers RJ, Kyba M, Ware CB, Nelson AM, et al. Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. PLoS Genetics. 2010;6(10):e1001181.

45. Miner BE, Stöger RJ, Burden AF, Laird CD, Hansen RS. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Research. 2004;32(17):e135.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

46. Shiraishi M, Hayatsu H. High-speed conversion of cytosine to uracil in bisulfite genomic sequencing analysis of DNA methylation. DNA Research. 2004;11(6):409–415.

47. DiCiccio TJ, Efron B. Bootstrap confidence intervals. Statistical science. 1996;p. 189–212.

48. Meeker WQ, Escobar LA. Teaching about approximate confidence regions based on maximum likelihood estimation. The American Statistician. 1995;49(1):48–53.

49. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015;520(7546):243–247.

Acknowledgments

We thank: three anonymous reviewers for insightful and constructive comments; Hehuang Xie, Ming-An Sun, and Julia Arand for generously sharing raw data from their earlier studies of methylation dynamics; Stanley Gartler, Barbara Wakimoto, Matthew Stephens, Elizabeth Thompson, Carl Bergstrom, Frank Stahl, Jette Foss, Takuo Yamaki, Jesse McClure, Audrey Mat, Hyun Ji Noh, Steve Henikoff, Angela Liang, Alice Tao, Brooks Miner, Simina Popa, Rachana Kumar, and M. Goe for insightful and helpful comments.

Supporting Information

Figure S1. Markov chain used for the derivation of RCP.

Figure S2. RCP in Dnmt3-knockout cells. (a) RCP values were inferred at the *Lep* locus for wildtype, *Dnmt3a*-KO, and *Dnmt3b*-KO murine ES cell lines, using data from Al-Azzawi *et al.* [30]. **(b-i)** RCP values were inferred for eight additional loci for wildtype, *Dnmt3a*-KO, *Dnmt3b*-KO, and *Dnmt3a/b*-double-KO murine ES cell lines, using data from Arand *et al.* [14].

Table S1. Comparison of RCP values inferred here to the DNMT1 hemi-preference ratios (HPR) inferred by Fu *et al.* [22]

Table S2. Hairpin-linkage and PCR conditions for collection of double-stranded DNA methylation patterns. Entries shown are for patterns published here for the first time. R.E. refers to the restriction enzyme used to create the genomic overhang prior to ligation with a hairpin linker.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S3. RCP values and associated confidence intervals inferred for the 24 data sets from our labs. We collected double-stranded DNA methylation patterns from two species — mouse and human — and several loci using bisulfite conversion under either low-molarity/temperature ("LowMT") or high-molarity/temperature ("HighMT") conditions [38]. For each data set, we counted methylated (M) , hemimethylated (H), and unmethylated (U) dyads, and used these values to infer methylation frequency, m, unmethylated dyad frequency, U, and the ratio of concordance preference, RCP. Inferences for (m, U) and for RCP incorporated correction for conversion error, as described in S3 Text. We estimated 95% confidence intervals by bootstrapping and applied the BCa correction to obtain bias-corrected confidence intervals and point estimates, as described in S7 Text. Both uncorrected and BCa-corrected intervals and point estimates are listed here.

Table S4. RCP values and associated confidence intervals inferred for data reported in Arand *et al.* [14, 17]. Dr. Julia Arand kindly shared raw double-stranded sequences for samples described in these publications. We accounted for failed and inappropriate conversion, as described in S3 Text, in our point estimates of m, U, and RCP. The inappropriate conversion rate of 0.031 was inferred because the bisulfite conversion conditions used by Arand et al. resembled lowMT conditions [38]. We estimated 95% confidence intervals by bootstrapping and applied the BCa correction to obtain bias-corrected confidence intervals and point estimates, as described in S7 Text. Both uncorrected and BCa-corrected intervals and point estimates are listed here. For one sample, Arand *et al.* (2015) *mSat* 3dpc, only dyad counts, but no raw sequences, were available. We used the dyad counts to estimate the point estimate and the confidence interval for this sample while assuming independent sampling of dyads (S8 Text).

Table S5. RCP values and associated confidence intervals inferred for data reported in Zhao *et al.* (2014) [15]. Dr. Hehuang Xie kindly shared raw data on M, H, and U values for samples described in Figure S6 of their paper, and also provided information on error rates from bisulfite conversion: for the "Day 0" sample (cultured stem cells grown under non-differentiating conditions), the failed conversion rate was 0.011, and the inappropriate conversion rate was 0.0109; for the "Day 6" sample (cultured cells grown for 6 days under differentiating conditions), the corresponding error rates were 0.012 and 0.0099. Dr. Xie also commented, "'All' refers to the total CG dyads and 'DNA' refers to the CG dyads within 'DNA repeat elements (DNA)' annotated in the UCSC genome database." For these data sets, whose individual sequence reads only contain 2 to 3 dyads on average, we assumed independent sampling of dyads to obtain the confidence intervals (S8 Text).

Table S6. Pairwise comparisons of RCP values inferred for replicate samples of cultured human ES and iPS cells, and for mouse embryonic cells sampled at various developmental stages. We applied our test for heterogeneity, as described in S5 Text, to assess evidence of significant differences between RCP values inferred from various sample replicates presented by Arand et al. (2015) [17]. A color gradient encodes approximate levels of significance, with dark green indicating non-significant differences, and red indicating highly significant differences.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

S1 Text: Deriving the Ratio of Concordance Preference, RCP

The goal of RCP is to quantify the extent to which the DNA methylation machineries that gave rise to each data set deviate from random expectations under the binomial distribution, as indicated by an over- or under-abundance of concordant dyads. Here we approach this problem by modeling the formation of concordant and discordant dyads as transitions between unmethylated and hemimethylated dyads and between hemimethylated and fully methylated dyads, without regard to the molecular processes that facilitate these transitions. We then take a mathematical approach to derive an expression for RCP.

We seek the equilibrium frequencies of fully methylated (M), hemimethylated (H), and unmethylated (U), dyads. Consider a continuous-time Markov chain operating on the probability distribution of the dyads $\langle M, H, U\rangle$:

$$\frac{d}{dt}\begin{pmatrix} M \\ H \\ U \end{pmatrix} = \begin{pmatrix} -2\gamma_1 & \eta_1 & 0 \\ 2\gamma_1 & -\eta_1 - \gamma_0 & 2\eta_0 \\ 0 & \gamma_0 & -2\eta_0 \end{pmatrix}\begin{pmatrix} M \\ H \\ U \end{pmatrix} \tag{2}$$

where η's and γ's represent the rates of methylation addition and removal, respectively, as shown in Fig S1.

In the meantime, we define RCP as the ratio between β_c, the rate of dyad conversions yielding concordant dyads, and β_d, the rate of dyad conversions yielding discordant dyads. We thus define $\beta_c = \sqrt{\eta_1\gamma_0}$, the geometric mean of the methyl-addition and methyl-removal rates yielding concordant dyads. Likewise, we define $\beta_d = \sqrt{\eta_0\gamma_1}$.

We can then solve for the steady state distribution for the Markov chain in Equation (2) to arrive at the ratio. We can also express in in terms of m and U, as shown below.

$$\text{RCP} := \frac{\beta_c}{\beta_d} = \frac{\sqrt{\eta_1\gamma_0}}{\sqrt{\eta_0\gamma_1}} = \sqrt{\frac{4MU}{H^2}} = \frac{\sqrt{U(U+2m-1)}}{1-U-m} \tag{3}$$

This formulation of RCP does not require the assumption that methylation frequency is constant over time. Here, "steady state" refers to the dyad frequencies expected under a given system of methylation processes, regardless of whether the methylation frequency is constant over rounds of cell division.

It is notable that RCP^2 is $4MU/H^2$, which is expected to equal 1 under the Hardy-Weinberg equilibrium [23], if dyad frequencies are considered as genotype frequencies of a gene with two alleles. Following this, RCP can be considered as a measure of deviation from the null equilibrium.

RCP is therefore a metric for the degree to which the system of methylation processes prefers concordance ($\text{RCP} > 1$), discordance ($\text{RCP} < 1$), or, possibly, exhibits no preference in either direction ($\text{RCP} = 1$). If we set $\text{RCP} = 1$ and solve this expression for U, we find that $U = (1-m)^2$. This is consistent with the expectation under the binomial distribution that RCP will be 1 when there is no preference for either concordance or discordance. As RCP approaches ∞, U approaches $1-m$. Setting $\text{RCP} = 0$ results in two solutions: $U = 1-2m$ and $U = 0$. These solutions are

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

congruent with the boundaries that define the space of (m, U) as given in "Defining the Ratio of Concordance Preference for All Possible Methylation Configurations" of the main text, and in Fig 1.

S2 Text: Comparing RCP Values and Hemi-Preference Ratios from HMM

"Hemi-preference ratio", a parameter inferred under our earlier analysis with a hidden-Markov model (HMM) [22], evaluates the preference of a given DNA methyltransferase for acting at hemimethylated as compared to unmethylated dyads [11], and thereby measures its preference for creating concordant dyads. Because RCP measures the concordance preference of the entire ensemble of enzymes that give rise to methylation patterns, the hemi-preference ratio of a given enzyme and the RCP inferred from the same data set are expected to have good agreement if that enzyme is the primary actor. The congruence between these two metrics is expected to decline with increasing contributions from other enzymes.

Three of the four data sets that had been analyzed under HMM showed very good agreement between the RCP values we infer here and the hemi-preference ratios previously inferred for the maintenance methyltransferase DNMT1: 58.0 vs. 58 for *FMR1*, 13.3 vs. 15 for *G6PD*, and 89.1 vs. 94 for *LEP* [22] (Table S1). The close correspondence between these values indicates that for these loci in leukocytes, methylation dynamics are driven primarily by conservative, maintenance-type processes such as accomplished by DNMT1, and that neither active demethylation nor *de novo* processes have a substantial role.

The fourth data set that had been analyzed under HMM, *LEP* in human adipose tissue, had an inferred RCP value of 34, well within the range of RCPs inferred for other data sets from single-copy loci including *LEP* in leukocytes (Fig 1a; Table S1). This RCP value was, however, more than eighteen-fold lower than the DNMT1 hemi-preference ratio estimate of 628 that we obtained under the earlier HMM approach (Table S1). What might account for this discrepancy? A hemi-preference ratio of 628 is unrealistically high, even for a maintenance enzyme, compared to the hemi-preference ratios inferred in other data sets, including those from methylation patterns established by DNMT1 *in vitro* [11]. This could reflect the inability of the HMM to yield a reasonable estimate for a data set impacted by demethylation, a process that was not considered in the HMM design. This lack of correspondence between HMM and RCP estimates is consistent with the possibility that active removal of methylation has a heightened role at loci with temporally variable transcription levels, and may reflect the role of the *LEP* locus as a sentinel of adipose but not blood [28].

S3 Text: Assessing the Potential Impact of Bisulfite-Conversion Errors

Two classes of error that occur during bisulfite-conversion can lead to misinterpretation of cytosine methylation states. Failure to convert unmethylated cytosines to uracil occurs at rate b, and can result in overestimation of methylation frequency. Inappropriate conversion of methylated cytosines to thymine, first noted by Shiraishi and Hayatsu, [46] occurs at rate c, and can result in underestimation of

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

methylation frequency. We have reported previously that the rates of failed and inappropriate conversion depend strongly on the chemical and thermal conditions of bisulfite conversion. [38] In particular, we found that bisulfite treatment prolonged beyond that required to attain complete or nearly complete conversion of unmethylated cytosines can yield high rates of inappropriate conversion. Historically, conversion protocols have been designed to minimize the failed-conversion rate, with little or no attention to the rate of inappropriate conversion events. Thus, while both classes of error can alter parameters used to infer RCP — the methylation frequency, m, and the unmethylated dyad frequency, U — errors arising through inappropriate conversion are of particular concern.

How severely can conversion error impact RCP? And to what extent does its potential impact depend on the true methylation frequency of the target sequences? Densely methylated sequences contain a large number of fully methylated dyads, such that the most likely conversion error is inappropriate conversion yielding apparent hemimethylated dyads. Such events elevate the apparent level of discordance in a given data set and artifactually depress the inferred RCP values. To quantify this potential impact, we calculated how RCP would be altered by a single artifactual hemimethylated dyad introduced by inappropriate conversion, assuming that in its true state the relevant data set had one unmethylated dyad, one hemimethylated dyad, and 98 methylated dyads. The inappropriate conversion of one cytosine among 98 dyads whose true state is fully methylated is close to the number expected for an inappropriate conversion rate of about 0.5%. For this hypothetical data set, the introduction of a lone artefactual hemimethylated dyad by an inappropriate conversion event reduces RCP from about 20 to about 10, i.e., by a factor of 2. Thus, in the absence of mathematical corrections of the sort implemented here, even very low levels of inappropriate conversion can severely impact inference of RCP.

Sparsely methylated sequences contain a large number of unmethylated cytosines that are potential targets for failed conversion. We calculated that for such sequences the impact of conversion error depends on whether most unmethylated cytosines are in hemimethylated or in unmethylated dyads. For example, when most unmethylated cytosines in a sparsely methylated sequence occur in hemimethylated dyads, the level of discordance is already high, such that the addition of an artifactual hemimethylated dyad by failed conversion only slightly decreases RCP. By contrast, when most unmethylated cytotosines are in unmethylated dyads, production of an artifactual hemimethyated dyad by failed conversion reduces RCP by a factor of two. Mathematical correction for conversion error is therefore critical, not only because error has potentially large impacts on absolute RCP values, but also because the magnitude of these impacts differs among data sets, with the potential either to magnify or to dampen variation among them.

S4 Text: Mathematical Correction for Bisulfite-Conversion Error

To account for conversion errors occurring at known rates, we express the observed dyad frequencies, M_{fobs}, H_{fobs}, and U_{fobs} , as functions of the true dyad frequencies M_t, H_t, and U_t, that would have been observed had conversion error not occurred:

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$
\begin{aligned}
M_{fobs}(M_t, H_t, U_t, b, c) &= M_t\,(1-c)^2 + H_t\,b\,(1-c) + U_t\,b^2 \\
H_{fobs}(M_t, H_t, U_t, b, c) &= 2\,M_t\,c\,(1-c) + H_t\,(1-b)\,(1-c) + 2\,U_t\,b\,(1-b) \\
U_{fobs}(M_t, H_t, U_t, b, c) &= M_t\,c^2 + H_t\,(1-b)c + U_t\,(1-b)^2
\end{aligned}
\tag{4}
$$

In cases where the mathematical correction yielded negative counts for one or two dyads, the negative counts were redistributed to the remaining dyads in proportion to their original counts, such that no dyad counts were negative after correction.

For each of the data sets reported here, DNAs were converted under one of two conditions: low molarity-temperature (LowMT), with failed-conversion rate 0.0030 and inappropriate conversion rate 0.031, and high molarity-temperature (HighMT), with failed-conversion rate 0.0086 and inappropriate-conversion rate 0.017 [38]. Error rates used in analysis of published data from other groups are given in Table S4 (Arand *et al.* [14, 17]) and Table S5 (Zhao *et al.* [15]). Using these rates and observed dyad frequencies, true dyad frequencies can be inferred by maximum likelihood.

S5 Text: Assessment of Heterogeneity Among Replicates and Quasi-Replicates

To ask about possible heterogeneity among RCP values inferred for replicates — for example, for individual developmental stages as investigated by Arand *et al.* [17] — and for quasi-replicates — for example, for multiple cell lines at similar cell-culture and differentiation conditions — we utilized pairwise, two-tailed permutation tests.

These calculations revealed an intriguing pattern. There was no evidence of heterogeneity among samples of cultured human ES and iPS cells ($p > 0.29$, pairwise two-tailed PTs for the six cultured stem cell lines at the *L1* locus; p-values summarized in Table S6; Fig 3). By contrast, several pairwise comparisons of mouse embryonic cells sampled independently for a given developmental stage yielded evidence of significant differences in RCP (p-values summarized in Table S6). Notably, however, most of the significant heterogeneity observed was for cells in pronuclear stages, at which data collection at identical stages is made difficult by rapid developmental transitions. On the other hand, there was evidence of only limited heterogeneity among totipotent and pluripotent cells between the 2-cell stage and the 3.5-dpc stage. Similarly, when experimental interventions were taken to prevent DNA replication or methylation (i.e. aphidicolin treatment at PN4/5 stages and SAMase treatment just after fertilization and before the first round of DNA replication), no evidence of heterogeneity was observed. Cell culture techniques are focused on minimizing opportunities for developmental differences to arise among cells, and could account for the observation of no heterogeneity among the cultured human ES and iPS cells.

Given the low levels of heterogeneity observed and the fact that some level of heterogeneity is inevitable in rapidly developing cells, we pooled several sets of replicates for analysis, as described in S6 Text (Fig 6).

S6 Text: Pooling Data Sets Across Replicates

In some instances, multiple independent replicates were collected and used to assess RCP for cells under a given biological condition — for example, multiple embryos at a

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

given pronuclear stage (Fig 6). In these cases, the sequence-level dyad counts of individual replicate data sets were corrected for conversion errors independently before the replicate data sets were pooled.

S7 Text: Inferring and Comparing RCP Without Assuming Independent Sampling of Dyads

CpG/CpG dyads typically are not sampled individually, but instead as members of sequence reads that can contain from a few to many neighboring dyads. Often, there is correlation among the methylation states of these neighboring dyads. The processivity of the DNA methyltransferases, especially DNMT1, is a substantial contributor to these correlations. As the mean number of dyads per read increases, so too does the potential for dyad-dyad correlation to undermine the accuracy of confidence intervals inferred under the assumption of independent sampling of dyads.

Of the three groups of data we analyzed here, one — Zhao *et al.* (2014) [15] — consists of reads that contain, on average, only two to three dyads, such that correlation among the dyads ascertained on a given molecule is anticipated to have a fairly minor impact on sampling. Reads for our data have as many as 22 dyads; mean dyad counts for data from Arand et al. [14, 17] are intermediate between our data and those of Zhao *et al.* (2014) [15].

For our own data and those of Arand *et al.* [14, 17], we used an approach that, at slightly higher computational cost, models and seeks to account for the potential impacts of dyad-dyad correlations. Our data yielded moderately larger confidence intervals under the bootstrapping approach as compared to under the likelihood approach with the assumption of independent sampling of dyads. By contrast, the data from Arand *et al.* [14, 17] yielded almost identical confidence intervals without and with the assumption of independence. In view of the even smaller mean number of dyads per read in the data of Zhao *et al.* (2014) [15], we chose to make the not unreasonable assumption that dyads were sampled independently in their data. Although it is possible that two or more reads originated from nearby regions of a single molecule and thus have dyad-dyad dependence, we assumed that the effect of such occurrences, if at all present, is very small, given the large amount of starting material.

The first of the two methods is described in this section, and the second in the next section, S8 Text.

Inferring RCP point estimates and confidence intervals. For our own data and those of Arand *et al.*, we used a bootstrapping approach to model the uncertainty in RCP values introduced by possible within-sequence correlations in methylation states, and to make point estimate and confidence-interval inferences that account for this uncertainty. A more rigorous examination of correctness for our bootstrapping approach is in process.

For each data set of n sequences, we sampled n sequences with replacement, $B =$ 2,000,000 times, and calculated the RCP for each of the B bootstrapped data sets. We inferred the true distribution of RCP for a given biological condition using the bootstrap distribution. It was clear from the resulting distributions that RCP is a biased estimator, as many of these distributions had longer right tails than left.

The simplest, and potentially misleading, approach for inference of point estimates

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and construction of confidence intervals from bootstrap distributions is to assume normality, and then to exclude right and left tails at the intended level of confidence. Efron and DiCiccio (1996) [47] commented that, for biased estimators, this approach can lead to inference of inappropriately exclusive limits at the long-tailed end of the distribution, and inappropriately inclusive limits at the right-tailed side.

To address this problem, we applied Efron and Diccicio's "bias-corrected and accelerated" (BCa) method. Under the BCa method, the cumulative-density function observed for the distribution of bootstrap replicates is compared to that expected under normality. Bias-corrected point estimates are then inferred as the 50th-percentile values in the BCa corrected distributions. Similarly, critical points for intervals of a given confidence level are inferred from the values at the relevant percentiles of the distribution of bootstrapped values.

Assessing whether a data set has RCP value greater than 1. If methyl groups are placed completely at random — that is, with preference for neither concordance or discordance — RCP is expected to be 1. In a previous report, Shipony *et al.* [16] interpreted their data to indicate that methyl groups are, indeed, placed essentially at random in undifferentiated cells. As there is very little evidence in any of our data sets to indicate possible preference for discordance, we opted to perform a one-tailed test, asking for the probability that our data sets do not have RCP values greater than 1.

To do so, we first calculated RCP point estimates for 200,000 bootstrap replicates, using the method described above, and then calculated the p-value as the fraction of those point estimates that were less than or equal to RCP of 1. For example, the finding that only 20 of the 200,000 bootstrap replicates yielded an RCP point estimate less than 1, we would conclude that, for this data set, RCP is significantly greater than 1 ($p = 0.0001$).

Assessing whether RCP values differ significantly between data sets. A key goal of our study is to assess possible evidence for RCP differences between data sets. For example, we ask whether RCP for a given cell type differs between samples collected under differentiating as compared to non-differentiating conditions.

To compute the significance of observed differences between RCP values for Data Sample A and Sample B, we used permutations to compute a null distribution of ordered differences (for example, RCP(Sample A) − RCP(Sample B)) expected under the null assumption that the sequences in the two sets were drawn from a single population. To do this, we pooled sequences from Sample A and Sample B and then repeatedly drew from that pool, without replacement, to generate, at random, versions of Sample A and Sample B with sequence counts equal to those of the observed data. For each pair of randomly generated sets, we calculated the difference between their RCP point estimates (θ^*) , and obtained the distribution of RCP-difference values under the null hypothesis. For one-tailed tests, for example, to ask whether one data set has RCP value greater than the RCP value for another, we took the proportion of the distribution greater than the difference of the point estimates ($\hat{\theta}$) as the p value. For two-tailed/equal-tailed tests, for example, to ask whether RCP values for two data sets differ significantly, we first calculated the proportion of permuted differences smaller than ($\hat{\theta}$). We then calculated the proportion of permuted differences greater than ($\hat{\theta}$). We took the twice the lesser of these two values as the p-value for observing a difference this great in the event that the sequences in the two data sets were, in reality, drawn from the same distribution.

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Defining the 95% confidence region for a data set in the (m, U) space, without assuming independent sampling of dyads. Using the methods described above, we generated 2,000,000 bootstrap samples of the data set. Instead of estimating the RCP value for each of the bootstrap samples, we calculated m and U. We constructed the two-dimensional confidence region for the two parameters for plotting using `ci2d` function in the `gplots` R package.

S8 Text: Inferring and Comparing RCP With Assuming Independent Sampling of Dyads

Calculating the likelihood of proposed true dyad frequencies, M_t, H_t, and U_t, given the observed dyad counts, $M_{c\,obs}$, $H_{c\,obs}$, and $U_{c\,obs}$. Failed- and inappropriate-conversion events create observed dyad frequencies that differ from true dyad frequencies. If the rates of these two types of error are known, the likelihood of a set of proposed true frequencies — M_t, H_t, and U_t — can be calculated as follows, given the observed dyad counts — $M_{c\,obs}$, $H_{c\,obs}$, and $U_{c\,obs}$ — and Equation (4). M_{fobs}, H_{fobs}, and U_{fobs}, which indicate the observed frequencies of the dyads, can be easily calculated from the dyad counts.

$$\begin{aligned}\mathcal{L}(M_t, H_t, U_t \mid M_{c\,obs}, H_{c\,obs}, U_{c\,obs}, b, c) = C &\times M_{fobs}(c, b, M_t, H_t, U_t)^{M_{c\,obs}} \\ &\times H_{fobs}(c, b, M_t, H_t, U_t)^{H_{c\,obs}} \\ &\times U_{fobs}(c, b, M_t, H_t, U_t)^{U_{c\,obs}} \\ \text{with multinomial coefficient } C = &\frac{(M_{c\,obs} + H_{c\,obs} + U_{c\,obs})!}{M_{c\,obs}!\; H_{c\,obs}!\; U_{c\,obs}!}\end{aligned} \tag{5}$$

Inferring RCP point estimates and confidence intervals. The RCP point estimate of a data set is calculated directly from the conversion-error-corrected observed dyad frequencies. We determine the 95% confidence interval for RCP as the interval that includes all values of RCP for which the natural log likelihood lies within $\chi^2_{0.95;\,df=1} = 1.92$ units of the maximum natural log likelihood point estimate. [48]

Although bias is just as much a concern with the assumption of independent sampling of dyads, we did not perform a bias correction for the data from Zhao *et al.* [15], because without bootstrapping, we lacked a simple method for bias estimation. Nonetheless, our observations from our analyses of other data sets suggest that most of these samples are likely not severely affected by bias. From bootstrapping of smaller data sets collected by our lab and by Arand *et al.* [14, 17], we have observed that the asymmetry in the distribution is small in the lower ranges of RCP (1∼10), but greater as RCP increases. Therefore, bias is likely to be small for most samples presented by Zhao *et al.*, for which RCP point estimates rarely exceed 10. For data sets presented by Zhao *et al.* that yielded RCP estimates greater than 10, bias are unlikely to affect our analyses and the general conclusion that their methylation behavior is characterized by strong preference for concordance.

Assessing whether RCP values differ significantly between data sets. To compare the RCP values between two data sets while assuming independence among all dyads, we can use a likelihood approach, comparing a model in which two true RCP values are required to described the two data sets to an alternate model in which both data sets can be explained by a single RCP value. We implemented that test as follows:

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Solving for U in Equation 3 gives:

$$U(m, \mathrm{RCP}) = 1 - m - \frac{1}{2}\left(\frac{1 - \sqrt{1 - 4m(1-m)\left(1 - \mathrm{RCP}^2\right)}}{1 - \mathrm{RCP}^2}\right) \tag{6}$$

Using this, we can also define $M(m, \mathrm{RCP})$ and $H(m, \mathrm{RCP})$. Modifying Equation 4, we can derive the expressions for $M_{fobs}(c, b, m, \mathrm{RCP})$, $H_{fobs}(c, b, m, \mathrm{RCP})$, and $U_{fobs}(c, b, m, \mathrm{RCP})$. We then can rewrite Equation 5, such that the parameters are c, b, m, and RCP:

$$\begin{aligned}\mathcal{L}_{one\ set}(M_{fobs}, H_{fobs}, U_{fobs} \mid c, b, m, \mathrm{RCP}) = C &\times M_{fobs}(c, b, m, \mathrm{RCP})^{M_{cobs}} \\ &\times H_{fobs}(c, b, m, \mathrm{RCP})^{H_{cobs}} \\ &\times U_{fobs}(c, b, m, \mathrm{RCP})^{U_{cobs}}\end{aligned} \tag{7}$$

We then employ a likelihood-ratio test to quantify the fit of an alternative model relative to the null. In the null model, which has three variable parameters, m^1_{null}, m^2_{null}, and RCP_{null}, one value of RCP explains both data sets. Using Equation 7:

$$\begin{aligned}&\mathcal{L}_{null}(M^1_{fobs}, H^1_{fobs}, U^1_{fobs}, M^2_{fobs}, H^2_{fobs}, U^2_{fobs} \mid c^1, b^1, m^1_{null}, c^2, b^2, m^2_{null}, \mathrm{RCP}) \\ &\qquad = L_{one\ set}(M^1_{fobs}, H^1_{fobs}, U^1_{fobs} \mid c^1, b^1, m^1_{null}, \mathrm{RCP}_{null}) \\ &\qquad \times L_{one\ set}(M^2_{fobs}, H^2_{fobs}, U^2_{fobs} \mid c^2, b^2, m^2_{null}, \mathrm{RCP}_{null}).\end{aligned} \tag{8}$$

The alternate model, which has variable four parameters, m^1_{alt}, m^2_{alt}, RCP^1_{alt}, and RCP^2_{alt}, has two values of RCP, one for each data set.

$$\begin{aligned}&\mathcal{L}_{alt}(M^1_{fobs}, H^1_{fobs}, U^1_{fobs}, M^2_{fobs}, H^2_{fobs}, U^2_{fobs} \mid c^1, b^1, m^1_{alt}, \mathrm{RCP}^1_{alt}, c^2, b^2, m^2_{alt}, \mathrm{RCP}^2_{alt}) \\ &\qquad = L_{one\ set}(M^1_{fobs}, H^1_{fobs}, U^1_{fobs} \mid c^1, b^1, m^1_{alt}, \mathrm{RCP}^1_{alt}) \\ &\qquad \times L_{one\ set}(M^2_{fobs}, H^2_{fobs}, U^2_{fobs} \mid c^2, b^2, m^2_{alt}, \mathrm{RCP}^2_{alt}).\end{aligned} \tag{9}$$

Computing the ratio of the maximum likelihoods for the null and alternate models, we can calculate the test statistic, D:

$$D = -2\, ln\left(\frac{\mathcal{L}_{null}(\hat{m}^1_{null}, \hat{m}^2_{null}, \hat{\mathrm{RCP}}_{null})}{\mathcal{L}_{alt}(\hat{m}^1_{alt}, \hat{\mathrm{RCP}}^1_{alt}, \hat{m}^2_{alt}, \hat{\mathrm{RCP}}^2_{alt})}\right) \tag{10}$$

Under the assumption of large sample of dyads, D is approximately χ^2 distributed with 1 degree of freedom.

Assessing whether a data set has RCP value greater than 1. We again take a likelihood-based approach as we did in the section above. Here, the null model states that the system operates under the specified RCP value. The alternate model states that the system operates under another RCP value.

$$\begin{aligned}&\mathcal{L}_{null}(\mathrm{RCP}_{\text{specified}}, m_1, ..., m_n \mid M_{c\ obs\ 1}, H_{c\ obs\ 1}, U_{c\ obs\ 1}, b_1, c_1, ..., M_{c\ obs\ n}, H_{c\ obs\ n}, U_{c\ obs\ n}, b_n, c_n) \\ &\qquad = \prod_i^n \mathcal{L}(m_i, \mathrm{RCP}_{\text{specified}} \mid M_{c\ obs\ i}, H_{c\ obs\ i}, U_{c\ obs\ i}, b_i, c_i)\end{aligned} \tag{11}$$

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\mathcal{L}_{alt}(\mathrm{RCP}_{alt}, m_1, ..., m_n \mid M_{c\ obs\ 1}, H_{c\ obs\ 1}, U_{c\ obs\ 1}, b_1, c_1, ..., M_{c\ obs\ n}, H_{c\ obs\ n}, U_{c\ obs\ n}, b_n, c_n) = \prod_i^n \mathcal{L}(m_i, \mathrm{RCP}_{alt} \mid M_{c\ obs\ i}, H_{c\ obs\ i}, U_{c\ obs\ i}, b_i, c_i) \tag{12}$$

We treat the two likelihood functions differently in maximizing them. For the alternate model, both n values of m and RCP_{alt} are parameters for maximization. For the null model, the RCP value is specified and thus fixed; only the n values of m are parameters for maximization.

$$D = -2\, ln \left(\frac{\mathcal{L}_{null}(\mathrm{RCP}_{\mathrm{specified}}, \hat{m_1}, ..., \hat{m_n} \mid ...)}{\mathcal{L}_{alt}(\hat{\mathrm{RCP}}_{alt}, \hat{m_1}, ..., \hat{m_n} \mid ...)} \right) \tag{13}$$

Under the assumption of a large sample of dyads, D is approximately χ^2 distributed with 1 degree of freedom.

Defining the 95% confidence region for a data set in the (m, U) space. We determine the 95% confidence region in the space of two parameters — here m and U — as the region that includes all proposed pairs of parameter values for which the natural log likelihood lies within $\chi^2_{0.95;\, df=2} = 3.00$ units of the maximum natural log likelihood point estimate [48].

S9 Text: Could Spontaneous Differentiation of a Subset of ES and iPS Cells Substantially Influence the Inference of RCP?

One possible explanation for the inference of conservative processes operating in cultured, undifferentiated cells is that these cells may in reality be a mixture of differentiated and undifferentiated cells. We calculated how large the differentiated subpopulation would need to be under this scenario, given that the subpopulation operated with the RCP inferred for the corresponding differentiated cells. The remainder of the population was assumed to operate at RCP of 1, per the alternate hypothesis. We allowed specification of m for each of the two populations.

Let p_1 and p_2 represent the proportions of the two putative subpopulations that we wish to estimate, such that $p_1 + p_2 = 1$. We start with RCP and m for each of the two subpopulations, which we denote by RCP_1, RCP_2, m_1, and m_2. We can find U_1 and U_2 using the RCP and m values. We have the following set of equations:

$$p_1 m_1 + p_2 m_2 = m_{\mathrm{overall}}$$
$$p_1 U_1 + p_2 U_2 = U_{\mathrm{overall}}$$

Using these equations and the observed overall RCP value, and the expression for RCP given m and U (Equation (3)), we can find p_1 and p_2.

Assuming that m of the undifferentiated subpopulation is at least 0.2, we found that over half of the cells would need to be differentiated if methylation in the remaining

bioRxiv preprint doi: https://doi.org/10.1101/072488; this version posted August 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

undifferentiated cells operates at RCP of 1. Morphological inspection of the cultured human stem cells did not suggest the presence of a substantial subpopulation of differentiated cells. Additionally, this calculation is not critical for the work presented here, because RCP values for totipotent and pluripotent stem cells from murine embryos are very similar to those inferred for all cultured ES and iPS cell lines that we examined. Such calculations may, however, be useful for future evaluation of cultured cells that are early in the process of differentiation.

S10 Text: Comparing RCPs of *Dnmt3*-Knockout Lines with Those of Wildtype Lines

The relative contributions of conservative processes are expected to be more substantial when the fraction of methylation events achieved through maintenance-type activity is elevated through loss of one or both of the *de novo* enzymes. Indeed, for the *Lep* locus, we inferred RCP values of 8.11 (*Dmnt3a* KO) and 6.97 (*Dnmt3b* KO), both higher than the RCP value for the wildtype murine ES cell line ($p = 0.008$, *Dnmt3a* KO; $p = 0.04$, *Dnmt3b* KO) (Fig S2a; Table S3).

Indeed, at the *Lep* locus, we inferred RCP values of 8.11 for *Dmnt3a* KO cells, a value significantly higher than that for wildtype cells ($p = 0.046$, two-tailed PT; Fig S2a; Table S3). Although the point estimate was higher for *Dmnt3b* KO cells as well, the difference was not significant ($p = 0.11$, two-tailed PT).

Examination of data from Arand *et al.* [14] for a broader set of loci in cell lines with knockouts at one or both of the Dnmt3s, however, revealed a more complex role for these enzymes in shaping methylation concordance. Overall methylation levels were diminished in the absence of the *de novo* methyltransferases 3a and 3b, as expected (Fig S2b-i, Table S3); RCP values increased as expected for several, though not all loci. Results for cells that lacked only one of the two DNA methyltransferases were even more variable across loci (Fig S2, Table S3). In some cases, loss of a single Dnmt3 enzyme had the predicted impact of increasing RCP (*L1* in 3a KO, $p < 10^{-6}$, Fig S2h; *L1* in 3b KO, $p = 0.002$, Fig S2h; *IAP* in 3a KO, $p = 0.002$, Fig S2g; *IAP* in 3b KO, $p = 0.029$, Fig S2g; two-tailed PTs), as was observed for *Lep* in our data. For one locus, *B1*, knockout of both of the Dnmt3s resulted in significantly increased RCP ($p < 10^{-6}$, two-tailed PT), though neither of the single knockouts did (Fig S2f). Most surprisingly, for two loci, loss of either of the Dnmt3s decreased RCP (*Igf2*, $p < 0.016$, Fig S2c; *Igf2*, $p < 10^{-6}$, Fig S2c; two-tailed PTs). These counterintuitive results likely reflect variation across loci in the roles of the individual DNA methlytransferases — and possibly the demethylation machinery — in shaping overall methylation levels across loci and categories of genic elements [49].