Title: A tissue specific atlas of gene promoter DNA methylation variability and the clinical value of its assessment

Authors: *Ryan H Miller[1], *Chad A Pollard[2], Kristin R Brogaard[1], Andrew C Olson[1], Larry I Lipshultz[5], Erica B Johnstone[6], Yetunde O Ibrahim[7], Jim M Hotaling[8], Enrique F Schisterman[3], Sunni L Mumford[4], *Kenneth I Aston[8], *Tim G Jenkins[2,8]

R Miller and C Pollard contributed equally

K Aston and T Jenkins contributed equally

Affiliations:

[1] Inherent Biosciences, Salt Lake City, UT, USA

[2] Department of Cell Biology and Physiology, Brigham Young University, Provo, UT, USA

[3] Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[4] Epidemiology Branch, Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Bethesda, Maryland

[5] Scott Department of Urology, Baylor College of Medicine, Houston, TX, USA

[6] Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, University of Utah School of Medicine, Salt Lake City, UT, USA

[7] Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and Gynecology, University of Texas Health Science Center in San Antonio, TX, USA.

[8] Division of Urology, Department of Surgery, University of Utah School of Medicine, Salt Lake City, UT, USA

*To whom correspondence should be addressed:

Tim Jenkins

Email: tim_jenkins@byu.edu

Abstract:

Background: Complex diseases have multifactorial etiologies making clinically actionable diagnostic markers difficult to identify. Novel tools with higher diagnostic yield and utility in driving personalized care are needed.

Methods: We utilized Illumina methylation array data from over 2400 samples to assess DNA methylation patterns in 20 distinct cell types ranging from sperm to brain as well as various disease states. We generated a simple analysis pipeline for DNA methylation data that focuses on intra-individual methylation variability within gene promoters. The analysis is designed, not to identify single causative gene alterations but instead focuses on any movement away from "healthy" methylation. This approach identifies altered regulation across multiple genes in related pathways thus enabling us to detect shifts in gene regulatory activity associated with distinct tissues and phenotypes. We explored three distinct questions in our assessment. 1) Are patterns of epigenetic variability tissue specific? 2) Do diseased tissues exhibit altered variability patterns compared to normal tissue? 3) Can epigenetic variability be detected in complex disease?

Results: Unsupervised clustering analyses established that patterns of epigenetic variability are tissue specific and that these patterns are at least as predictive of tissue type as differential methylation analysis. We demonstrated the ability to use these patterns to differentiate between healthy and diseased tissue with unsupervised clustering even in cases of complex

multifactorial diseases. We applied this method to the clinical use case of male infertility and found that men undergoing intrauterine insemination (IUI) with the lowest number of epigenetically dysregulated promoters in their sperm were almost twice as likely to father a child than men participating in IUI with the highest number of dysregulated promoters (p=0.011). We saw no significant difference in birth rates between groups of men with high and low numbers of dysregulated promoters undergoing in vitro fertilization (IVF), indicating IVF as a better treatment than IUI to achieve live birth in the presence of multi-pathway dysregulation in sperm.

Conclusions: This study demonstrates that patterns of epigenetic variability can differentiate between tissue types. While intuitive, this finding has never been demonstrated previously and suggests that specific epigenetic variability patterns may be used to predict phenotypic changes in disease states as these are, by definition, functional changes to cellular phenotypes. We demonstrate that the variability of gene regulatory marks are distinct between healthy and diseased tissue. This is particularly apparent at genes known to be important to cell function of the tissue of interest. While in some cases these regional alterations can be seen across the entire genome, more often the regulatory alterations that define a pathological phenotype are restricted to genes of known importance to a particular tissue. Importantly, in the case of sperm, we found that these patterns of variability did have utility in predicting infertile patients who would conceive through intrauterine insemination (IUI). We would propose that this discriminatory ability is due to the fact that the signature can be assessed in an n-of-1 context and that the patterns of variability identify any shift away from regulatory normalcy in pathways known to be impactful in the tissue of interest, and not only assessing the presence or absence of rare genetic variants. While the data presented here are encouraging, more work needs to be performed in other tissues to determine when, and in what context, these findings could be clinically actionable.

78 **Introduction:**

79 In 2003, one of the most profound efforts ever undertaken in the biological sciences, the Human
80 Genome Project, was completed[1]. At the time there was a great deal of hope that unlocking
81 the genetic code was the key to diagnosing and treating the vast majority of diseases. While the
82 discoveries made have been of great interest to many and have opened the door for important
83 genetic and epigenetic findings, clinically meaningful impacts remain elusive for many diseases.
84 In large part, this is due to the complex, multifactorial nature of most disease processes, with
85 etiologies resulting from a constellation of genetic, epigenetic, and environmental perturbations.
86 As a result, novel approaches to our analysis of genetic and epigenetic data are needed to
87 identify clinically actionable predictors of disease or disease progression.

88

89 In hindsight, the lack of clinical efficacy for many epigenetic and genetic findings is not
90 surprising. While it is common to identify a genetic or epigenetic association to a given disease,
91 those associations occur so rarely in a population, or are only one of many alterations required
92 to generate a pathological phenotype, that it is unreasonable to produce a highly predictive
93 screening test capable of impacting clinical care. Importantly, while the causative factors of
94 complex disease are multifactorial and highly variable, the tissue and cellular phenotypes that
95 occur as a result of the disease are likely to be more conserved. Therefore, novel diagnostic
96 approaches should focus, not on single genes or independent modifications associated with a
97 pathology, but on a holistic screen of alterations to gene regulatory activity at genes specifically
98 important to the affected tissue.

99

100 Certain data types are ideal for analyses focused on perturbations to gene regulatory networks.
101 DNA methylation is of particular interest in this effort. Because a tissue's phenotype is defined
102 by gene activity, and gene activity is controlled (at least in part) by epigenetic marks, it is
103 intuitive that marks such as DNA methylation are effectively the fingerprint of cell and tissue

type. A logical extension of this idea is the fact that DNA methylation and other epigenetic marks are altered in diseases such as cancer[2,3] and type 2 diabetes[4,5] not because they are causative features of the disease, but because the diseases induce epigenetic modifications to achieve a perturbed phenotype. Further, unlike the fairly static nature of the genome, DNA methylation is a dynamic biomarker affected by a host of factors such as age[6,7] and various modifiers including obesity[8,9], exercise[10,11], and environment[12]. Despite the evidence that clearly suggests that DNA methylation is tied to complex disease states, we need to use novel analytical approaches to realize its full potential.

Herein, we present a novel approach to DNA methylation analysis and an assessment of its efficacy as a clinical predictor. One recent study used a similar method specifically for women's cancers and found significant utility of the approach[13]. In our case, we describe a simple analysis of promoter DNA methylation variability that allows us to assess gene regulatory networks in a novel way. We provide evidence using DNA methylation array data from over 2400 samples and 20 cell types that our analysis can be used to identify and assess the genes most tightly regulated in specific cell types and that these patterns are highly tissue-specific. We then show the utility of this approach in an n-of-1 analysis to demonstrate that there are gene regulatory network perturbations common among individuals who suffer from specific pathologies in many different tissues. Lastly, we demonstrate in sperm the potential clinical value of this approach.

Methods:

Data Collection

Several publicly available datasets were used in this study. Infinium HumanMethylation450 Bead Chip data was obtained for tumor and healthy tissue samples (n=494) from The Cancer Genome Atlas (TCGA) Program as compiled by the University of California Santa Cruz Xena

130 Functional Genomics Explorer[14] (https://xenabrowser.net/datapages/). Infinium
131 HumanMethylation450 Bead Chip data for CD4+ T cell (n=11), CD8+ T cell (n=14), Alzheimer's
132 disease and control brain (n=190), lung (n=6), liver (n=26), and skin methylation data (n=18)
133 from healthy and diseased individuals were accessed from the NIH Gene Expression Omnibus
134 (GSE130029, GSE130030, GSE66351, GSE51077, GSE61258, GSE115797, respectively).
135
136 Sperm Infinium HumanMethylation450 Bead Chip data from fertile sperm donors as well as
137 patients undergoing in vitro fertilization (IVF) (n=166) was used from a previously published
138 single-site study by Aston, et al[15] as well the sperm Infinium MethylationEPIC Array data from
139 a clinical multi-site study of patients being seen by physicians for fertility care (n=1311) as
140 published by Jenkins et al[16].
141
142 *Sample Collection*
143 Semen samples were procured from University of Utah Andrology department from consented
144 patients undergoing fertility care (n=60), as well as two independent fertile sperm donor cohorts
145 (n=59). Semen samples from consented patients seeking fertility care were also procured from
146 the Urology Department at Baylor College of Medicine (n=49).
147
148 *Sample Preparation*
149 For all semen samples, somatic cell lysis, sperm isolation, DNA extraction, and bisulfite
150 conversion were performed as described by Aston, et al[15]. The bisulfite converted sperm DNA
151 was hybridized to Illumina Infinium HumanMethylationEPIC microarrays and ran as
152 recommended by the manufacturer (Illumina) at Infinity BiologiX.
153
154 *Data Preprocessing*

155 Figure 1 contains a flow chart of data processing and statistical analysis. The raw methylation
156 array data from the sperm, neuron, glia, skin, CD4+ T cell, and CD8+ T cell samples were
157 preprocessed using the minfi R package[17] using SWAN normalization to produce beta and m-
158 values for each cytosine-guanine dinucleotide (CpG). Density plots of the beta values of each
159 sample were examined to ensure the distribution of beta values followed a bimodal distribution
160 with prominent peaks between 0.0-0.2 and 0.8-1.0 and flat valleys from 0.2-0.8. Any samples
161 not following this distribution were removed and the remaining samples were renormalized. Beta
162 values are described as (methylated probe intensity / [methylated + unmethylated probe
163 intensity + 100]) and range from 0-1 with values around 0 being unmethylated and values
164 around 1 being methylated. M-values are described as (log(methylated probe intensity /
165 unmethylated probe intensity) and are useful measures of methylation to prevent bias arising
166 from heteroscedasticity seen when analyzing beta values[18] (Supplementary Figure S1,
167 Additional File 1).
168
169 For data processing of sperm samples, we removed any sperm samples from analysis that did
170 not have a mean methylation value less than 0.20 of all the CpG beta values in the differentially
171 methylated region of DLK1 as described by Jenkins, et al[19] (chr14:101,191,893-101,192,913,
172 GRCh37). According to Jenkins et al, the methylation states of the probes in this region are a
173 good discriminator between sperm and somatic cells and this procedure ensured analyses were
174 only performed on samples containing sperm DNA methylation and not contaminating somatic
175 cell DNA methylation.
176
177 Raw data for the TCGA datasets as assembled on the UCSC Xena platform and the lung and
178 liver datasets (GSE51077 and GSE61258, respectively) were not available, so the available
179 beta values were used. These beta values were logit-transformed to obtain the m-values for
180 these samples.

181

182 *Statistical analysis*

183 We define a given gene promoter as the genomic region one kilobase upstream and one
184 kilobase downstream from the transcription start site of a given gene. A gene promoter needed
185 to contain five or more methylation array probes to be used in any downstream analysis. Gene
186 methylation promoter variability (or "promoter variability") is defined as the standard deviation of
187 the m-values of the methylation array probes present in a defined promoter region
188 (Supplementary Figure S2, Additional File 1).

189

190 Hierarchical clustering was performed on all promoter variability values of samples from various
191 tissue types using the R software package 'pheatmap' (R version 4.0.3) with default parameters.
192 In cases where more than 20 samples existed for a given tissue, 20 samples were randomly
193 selected for inclusion in the clustering analysis to give a more uniform number of samples per
194 tissue type. Principal component analyses were performed on all promoter variability values
195 using the 'sklearn' library in Python (Python version 3.7.3).

196

197 We found the most epigenetically stable promoters of a given tissue type by identifying the
198 promoters with the lowest levels of variability in healthy samples of that tissue type. We did this
199 by first calculating a stability threshold for each promoter in a given tissue. A promoter stability
200 threshold represents the highest level of variability we expect to see in a given promoter of a
201 healthy sample of a given tissue (Supplementary Figure S2, Additional File 1). Then, the
202 promoters were rank ordered by the stability threshold values in ascending order. For the
203 analyses comparing promoters across tissue types (Figures 2B, 2C, 2F, 2G), we defined the
204 most stable promoters as the top first percentile of promoters with the lowest stability thresholds
205 in healthy samples of the given tissue. The most stable promoters for the sperm n-of-1 analyses

206 were defined as the top 10th percentile of promoters with the lowest stability thresholds in fertile
207 sperm donors.
208
209 Sperm n-of-1 analyses were performed by finding the most stable promoters in a cohort of fertile
210 sperm donors and counting the number of dysregulated promoters in each sample. We defined
211 a dysregulated promoter as a promoter that fell above the corresponding variability threshold.
212 Samples with the lowest number of dysregulated promoters are most similar to healthy controls.
213
214 Statistical differences in the pregnancy and live birth rates of men undergoing intrauterine
215 insemination (IUI) and in vitro fertilization (IVF) with the least and most dysregulated promoters
216 were calculated with two-sided t-tests. The men undergoing IUI had been through an average of
217 2.5 IUI attempts.
218
219 Tissue-specific gene ontology enrichment analyses were performed by running the PANTHER
220 Overrepresentation Test (http://pantherdb.org/webservices/go/overrep.jsp) on the gene names
221 of the first percentile of most stable promoters in a given tissue. Each test was run using a
222 background gene set that consisted of all genes with promoters containing five or more
223 methylation array probes.
224
225 **Results:**
226 *Tissues have unique methylation variability signatures*
227 Using microarray DNA methylation data, we explored the differences in gene promoter
228 methylation variability of various healthy tissues. Unsupervised clustering of all gene promoter
229 variability values showed tissue specificity and also revealed similarities among related tissues,
230 Figure 2A. For example, gastrointestinal tissue samples such as those from the esophagus,
231 stomach, colon and rectum, clustered closely together. We likewise saw clustering of samples

232 associated with the immune system (CD4+ T cells, CD8+ T cells, thymus), female reproduction
233 (endometrium, cervix), and brain (glia, neurons).
234
235 *Tissues are regulated at tissue-specific biological pathways*
236 We sought to identify how promoter variability differs among various tissue types. We identified
237 the most stable promoters in sperm and assessed the average methylation variability values for
238 these promoters in many samples across several tissue types as seen in Figure 2B. At
239 promoters indicated as most stable in the male germline, sperm samples have significantly
240 lower mean values than other tissue samples. Gene ontology analysis of these sperm
241 promoters show significant enrichment for sperm-related biological processes (Figure 2F). We
242 also looked at the mean of the promoter variability values of the known sperm-related genes
243 protamine 1 (PRM1), protamine 2 (PRM2), and protamine 3 (PRM3) which are genes expressed
244 exclusively in sperm and replace the majority of histones to achieve extreme nuclear
245 compaction in this specialized cell[20]. As expected, sperm samples displayed significantly less
246 variability in these promoters than other tissues (Figure 2D). These same analyses were
247 performed for the most stable promoters in neurons (Figures 2C, 2E) and a known neuron-
248 specific gene, CASP8 (Figure 2G) with similar results. Additional File 1 (Supplementary Figure
249 S4, Additional File 1) contains the results of these same analyses performed for several other
250 tissue types. It is important to note that while the most stable promoters in a given tissue are
251 generally characterized by very low promoter variability in samples from the given tissue, these
252 promoters all have varying degrees of absolute methylation (hypo, mid, or hyper-methylation), a
253 feature missed by traditional differentially methylated region (DMR) analyses (Supplementary
254 Figure S5, Additional File 1). It is also conceivable that this method can help overcome technical
255 biases inherent to methylation microarrays such as batch effects.
256
257 *Methylation variability can differentiate between healthy and diseased tissue*

258 In addition to distinguishing between tissue types, analysis of promoter methylation variability
259 can enable the differentiation of diseased and healthy tissue samples of the same tissue type.
260 One notable example is the ability to distinguish between tumor and healthy tissue based on
261 promoter variability signatures. Figure 3A depicts the first two principal components of promoter
262 variability values for colon primary tumor tissue and healthy colon tissue. The healthy colon
263 tissue samples appear to be tightly clustered together, whereas the tumor samples are widely
264 distributed throughout the plot. Figure 3B shows the difference in promoter variability between
265 psoriatic skin lesions and adjacent healthy skin samples from the same individuals. Figure 3C
266 shows a principal component analysis of neurons, glial cells, as well as bulk cell samples from
267 postmortem brain tissue of individuals with Alzheimer's disease and controls. The plot shows
268 clear separation among the neurons, glial cells, and bulk cell samples indicating a difference in
269 promoter variability among different cell types in the same tissue. There is also separation
270 between control and Alzheimer's disease samples in neuron and glial cell samples but such
271 separation is not apparent among the bulk cell samples which may suggest subtle differences in
272 promoter variability might be more apparent when samples are sorted for individual cell types.
273
274 *Methylation variability of sperm can identify a subset of men with male factor infertility*
275 We performed n-of-1 analyses on over 1500 sperm samples from fertile sperm donors as well
276 as men being treated for male factor infertility. Figure 4A shows there was a significantly higher
277 number of dysregulated promoters in men being treated for male factor infertility compared to
278 fertile sperm donors. Supplementary Figure S8 (Additional File 1) shows the number of
279 dysregulated promoters in multiple cohorts of sperm samples, including the sperm donor cohort
280 used to find the most stable promoters in sperm as well as the stability thresholds. To give a
281 visual explanation of promoter methylation, Figure 4B depicts the promoter variability values
282 (red dots) of the most stable sperm promoters and the corresponding stability thresholds for
283 these promoters (black line) in an individual fertile sperm donor sample as well as a patient

284 being treated for male factor infertility (Figure 4C). It is clear that the patient being treated for
285 male factor infertility has many more dysregulated promoters than the fertile sperm donor. This
286 suggests that male factor infertility may be more related to a global shift in methylation variability
287 at promoters important for sperm cells rather than single nucleotide changes or epimutations.
288
289 We then looked at the relationship between dysregulated promoters and pregnancy and live
290 pregnancy rates in 1311 individuals being seen by a physician for infertility care. Table 1A
291 shows that in men undergoing intrauterine insemination (IUI), those with least number of
292 dysregulated promoters (lowest 10th percentile of patients) had significantly higher pregnancy
293 and live birth rates than IUI patients with the most number of dysregulated promoters (top 10th
294 percentile of patients). However, we saw no difference in pregnancy and live birth rates when
295 comparing men undergoing in vitro fertilization (IVF) with the least and most number of
296 dysregulated promoters, suggesting IVF should be the preferred fertility treatment option for
297 men with high levels of methylation dysregulation.
298
299 **Discussion:**
300 Here, we introduce a novel method which assesses gene promoter DNA methylation variability
301 to identify highly regulated genes in multiple tissue types and how these can be impacted in
302 various disease states. Hierarchical clustering of gene promoter variability from many tissues
303 demonstrates how unique these patterns are in different tissues and how these patterns remain
304 largely consistent in related, but distinct tissues. For example, numerous tissues from the
305 gastrointestinal tract cluster together as do tissues important to the function of the immune
306 system.
307
308 We also found that the most stable promoters in a given tissue have significantly lower
309 methylation variability than the same promoters in other tissues highlighting the importance of

310 unique genes and gene networks to any given tissue's function. Importantly, when assessing
311 DNA methylation variability within the same tissue type, we were also able to visualize
312 differences between healthy and diseased tissues such as in cancer, psoriasis, and Alzheimer's
313 disease.
314
315 To highlight the potential clinical impact of the assessment of promoter level DNA methylation
316 variability, we examined the pattern's utility in an assessment of male factor infertility and found
317 that men being seen by a physician for infertility had much higher levels of dysregulated
318 promoters. In addition, men undergoing IUI treatments with the most number of dysregulated
319 promoters had significantly lower pregnancy and live birth rates compared to men undergoing
320 IUI treatments with the least number of dysregulated promoters. However, this stark difference
321 in pregnancy and live birth rates was not seen between men with the highest and lowest levels
322 of dysregulated promoters in men undergoing IVF. This finding has great clinical significance
323 because it suggests that if a man is struggling with infertility and has a high level of promoter
324 dysregulation, he has much better odds at having a child after undergoing IVF than if he simply
325 went through multiple rounds of IUI.
326
327 Taken together, these data suggest that promoter DNA methylation variability is an excellent
328 indicator of tissue type. This is likely due to the fact that the analysis of variability is able to
329 successfully detect genes that are the most tightly regulated (via DNA methylation in this case)
330 in any given tissue. The assumption is that these genes must play an important role in cell
331 function unique to each tissue. This is supported by our data that demonstrated that promoter
332 DNA methylation variability is increased on average in abnormal tissues when compared to
333 normal tissue. This was particularly apparent in our assessment of sperm DNA methylation
334 variability patterns. While much work still needs to be performed, these patterns are promising
335 and may prove to have significant utility in clinical decision making.

336

337 Of additional value in our assessment of this approach and its translational capacity is the ability
338 to perform these analyses in an n-of-1 context. Specifically, because we are assessing intra-
339 promoter variability within a single individual, we are able to reliably assess variability in a single
340 individual with limited concerns of batch effects which would require normalization. While this
341 has yet to be fully vetted in the current work, it does appear that these analyses, if proven to
342 have high predictive power, could be translated to a diagnostic tool.

343

344 While we did perform one of the largest analyses to date in terms of tissue types and sample
345 numbers, many questions still remain to be answered. Among the most critical is the utility of
346 this analysis in many different tissues and disease states. Because we were only able to
347 perform a deep analysis in sperm due to the large numbers of samples, similar work needs to
348 be done in other tissues to determine if this analysis can provide clinically actionable
349 information. Even in our sperm study, this work needs to be replicated in independent cohorts to
350 determine its efficacy. One area of concern for this type of analysis is the fact that not all tissues
351 are able to be as easily purified as sperm. In fact, we found that when looking at bulk tissues in
352 the brain, we lost any ability to discriminate between healthy and diseased tissue. Thus, future
353 work needs to focus on as pure of tissue as possible to yield meaningful results. This could
354 create technical challenges that make the utilization of this analysis untenable in some
355 applications. Despite this, there will certainly be tissues and diseases in which the assessment
356 of promoter DNA methylation variability will be feasible.

357

358 These findings provide a novel means to define which genes each cell and tissue type tightly
359 regulate to ensure their unique phenotype and function. Because these signals have potential
360 utility in both the a basic understanding of tissue specific epigenetic patterns and in the clinical
361 assessment of diseased tissues, as well as the prediction of outcomes, this work provides

important foundational findings upon which tissue and disease specific assessments can be constructed in the future. While much work remains to determine clinical actionability for various applications, the results here are encouraging and may offer another tool with which we can assess the health of tissues and, importantly, predict the outcomes from various clinical interventions.

Availability of data and materials:

The data from this study are available from the corresponding author upon reasonable request.

References:

1. Collins FS, Morgan M, Patrinos A. The Human Genome Project: Lessons from Large-Scale Biology. Science [Internet]. American Association for the Advancement of Science; 2003 [cited 2022 Feb 3]; Available from: https://www.science.org/doi/abs/10.1126/science.1084564

2. Zhao SG, Chen WS, Li H, Foye A, Zhang M, Sjöström M, et al. The DNA methylation landscape of advanced prostate cancer. Nat Genet. 2020;52:778–89.

3. Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. Trends Genet. Elsevier; 2021;37:1012–27.

4. Davegårdh C, García-Calzón S, Bacos K, Ling C. DNA methylation in the pathogenesis of type 2 diabetes in humans. Mol Metab. 2018;14:12–25.

5. Ahmed SAH, Ansari SA, Mensah-Brown EPK, Emerald BS. The role of DNA methylation in the pathogenesis of type 2 diabetes mellitus. Clin Epigenetics. BioMed Central; 2020;12:1–23.

6. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. BioMed Central; 2013;14:1–20.

7. Jenkins TG, Aston KI, Cairns B, Smith A, Carrell DT. Paternal germ line aging: DNA methylation age prediction from human sperm. BMC Genomics. BioMed Central; 2018;19:1–10.

8. Benton MC, Johnstone A, Eccles D, Harmon B, Hayes MT, Lea RA, et al. An analysis of DNA

15

387 methylation in human adipose tissue reveals differential modification of obesity genes before
388 and after gastric bypass and weight loss. Genome Biol. BioMed Central; 2015;16:1–21.
389 9. Godfrey KM, Sheppard A, Gluckman PD, Lillycrop KA, Burdge GC, McLean C, et al.
390 Epigenetic Gene Promoter Methylation at Birth Is Associated With Child's Later Adiposity.
391 Diabetes. American Diabetes Association; 2011;60:1528–34.
392 10. Barrès R, Yan J, Egan B, Treebak JT, Rasmussen M, Fritz T, et al. Acute Exercise
393 Remodels Promoter Methylation in Human Skeletal Muscle. Cell Metab. Elsevier; 2012;15:405–
394 11.
395 11. Rönn T, Volkov P, Davegårdh C, Dayeh T, Hall E, Olsson AH, et al. A Six Months Exercise
396 Intervention Influences the Genome-wide DNA Methylation Pattern in Human Adipose Tissue.
397 PLOS Genet. Public Library of Science; 2013;9:e1003572.
398 12. Tobi EW, Lumey LH, Talens RP, Kremer D, Putter H, Stein AD, et al. DNA methylation
399 differences after exposure to prenatal famine are common and timing- and sex-specific. Hum
400 Mol Genet. 2009;18:4046–53.
401 13. Bartlett TE, Jones A, Goode EL, Fridley BL, Cunningham JM, Berns EMJJ, et al. Intra-Gene
402 DNA Methylation Variability Is a Clinically Independent Prognostic Marker in Women's Cancers.
403 PLOS ONE. Public Library of Science; 2015;10:e0143178.
404 14. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and
405 interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38:675–8.
406 15. Aston KI, Uren PJ, Jenkins TG, Horsager A, Cairns BR, Smith AD, et al. Aberrant sperm
407 DNA methylation predicts male fertility status and embryo quality. Fertil Steril. 2015;104:1388-
408 1397.e5.
409 16. Jenkins T, Aston K, Carrell D, DeVilbiss E, Sjaarda L, Perkins N, et al. The impact of zinc
410 and folic acid supplementation on sperm DNA methylation: results from the folic acid and zinc
411 supplementation randomized clinical trial (FAZST). Fertil Steril [Internet]. 2021 [cited 2021 Dec
412 1]; Available from: https://doi.org/10.1016/j.fertnstert.2021.09.009

413 17. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al.
414 Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA
415 methylation microarrays. Bioinformatics. 2014;30:1363–9.
416 18. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and
417 M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics.
418 BioMed Central; 2010;11:1–9.
419 19. Jenkins TG, Liu L, Aston KI, Carrell DT. Pre-screening method for somatic cell
420 contamination in human sperm epigenetic studies. Syst Biol Reprod Med. Taylor & Francis;
421 2018;64:146–55.
422 20. Balhorn R. The protamine family of sperm nuclear proteins. Genome Biol. BioMed Central;
423 2007;8:1–8.
424

425 **Figure Legends:**

426 *Figure 1: Data processing and statistical analysis workflow.*
427 Processing and analysis of Illumina's Infinium HumanMethylation450 and
428 HumanMethylationEPIC array data from multiple tissue types to derive promoter variability and
429 promoter stability thresholds and analyze their relationships among tissue types and between
430 healthy and diseased tissues.
431

432 *Figure 2: Tissues have unique patterns of gene methylation promoter variability*
433 **A)** Hierarchical clustering of promoter variability of 18 sample types representing five biological
434 systems. **B)** and **C)** Average promoter variability of 6 distinct cell types in the most stable
435 promoters (top 1st percentile) in sperm and neurons, respectively. One dot represents one
436 sample, and boxplots are overlaid to show the distribution of average promoter variability of

437 each tissue. All p-values comparing methylation variance between sperm and neuron to other
438 tissues types were ≤ 5.16E-14. **D)** Average protomer variability in the 6 cell types of three
439 sperm-specific protamine promoters. **E)** Promoter variability from one neuron specific apoptosis
440 promoter in the 6 cell types. All p-values for panels D and E when comparing sperm and
441 neurons to other cell types were ≤ 9.99E-17. **F)** and **G)** Gene ontology enrichment of the most
442 stable promoters for sperm and neurons, respectively.
443
444 *Figure 3: Diseased tissue samples have unique patterns of gene methylation promoter*
445 *variability compared to healthy*
446 **A)** Principal component analysis of promoter variability values from primary colon tumors and
447 normal colon tissue. **B)** Principal component analysis of promoter variability values from
448 matched psoriatic lesion and healthy skin samples. **C)** Principal component analysis of promoter
449 variability among neurons, glial cells, and bulk cells from postmortem brains of individuals with
450 Alzheimer's disease as well as controls.
451
452 *Figure 4. Dysregulated promoters are enriched in men seeking fertility care compared to fertile*
453 *controls*
454 **A)** Dysregulated promoters in samples from five independent studies. The most stable sperm
455 promoters and corresponding stability thresholds were calculated from a cohort of fertile sperm
456 donor samples. Circle colors represent the following: blue - Jenkins et al[16], green - Baylor
457 College of Medicine, purple - fertile sperm donors, Red and Orange - University of Utah
458 Andrology two independent collections. **B)** Promoter variability at the most stable sperm
459 promoters in a single fertile donor sperm sample (red dots). The stability threshold for these
460 promoters are shown in black. A red dot above the black line indicates a dysregulated promoter.
461 **C)** Same analysis as in B) but for a single patient being treated for male factor infertility.

462

463 *Table 1. Pregnancy and live birth rates in patients with the least and most dysregulated*

464 *promoters*

465 **A)** Pregnancy and live birth rates from male patients undergoing on average 2.5 intrauterine

466 insemination (IUI) cycles (N=528). **B)** Pregnancy and live birth rates from male patients

467 undergoing IVF (N=234). For each patient's sperm sample, the number of dysregulated

468 promoters were counted. Pregnancy and live birth rates were compared between the patient

469 cohort with the top 10th percentile bottom 10th percentile of dysregulated promoters.

470

471 **Supplementary Information Figure Legends:**

472 *Additional File 1: Supplementary figures and tables*

473 **Figure S1**. Heteroscedasticity of beta values in a sperm donor sample. **Figure S2**. Variability

474 equations. **Figure S3**. Hierarchical clustering of different control tissues using beta values

475 **Figure S4**. Average promoter variability at tissue-specific promoters. **Figure S5**. Example of

476 promoters with low methylation variability but varying levels of methylation. **Figure S6**. Principal

477 component analysis of diseased and control samples. **Figure S7**. Hierarchical clustering of

478 diseased and control tissue samples. **Figure S8**. N-of-1 dysregulated promoter analysis of

479 sperm samples from multiple cohorts. **Table S1**. Pregnancy/Birth rates in men with normal

480 sperm concentration (≥ 15 million sperm / mL). **Table S2**. Statistics from men with the least and

481 most number of dysregulated promoters. **Table S3**. Ranking by sperm concentration.

482

483 **Acknowledgements:**

484 The results shown here are in part based on data generated by the TCGA Research Network:

485 https://www.cancer.gov/tcga.

Funding:

The Folic Acid and Zinc Supplementation Trial was supported by contracts HHSN275201200007C and HHSN275201300026I from the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development. SLM was supported in part by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland. The work focused on sperm dysregulation was supported in part by Inherent Bioscience's National Science Foundation The Small Business Innovation Research (SBIR) Grant No. (2034014).

Author information:

Ryan H Miller and Chad Pollard contributed equally to this work. Tim G Jenkins and Kenneth I Aston contributed equally to this work.

Affiliations:

Inherent Biosciences, Salt Lake City, UT, USA

Ryan H Miller, Kristin R Brogaard, Andrew C Olson, Tim G Jenkins

Department of Cell Biology and Physiology, Brigham Young University, Provo, UT, USA

Chad Pollard, Tim G Jenkins

Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

20

510 Enrique Schisterman

511

512 *Epidemiology Branch, Division of Intramural Population Health Research, Eunice Kennedy*

513 *Shriver National Institute of Child Health and Human Development, Bethesda, Maryland*

514 Sunni L Mumford

515

516 *Scott Department of Urology, Baylor College of Medicine, Houston, TX, USA*

517 Larry Lipshultz

518

519 *Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and*

520 *Gynecology, University of Utah School of Medicine, Salt Lake City, UT, USA*

521 Erica B Johnstone

522

523 *Division of Reproductive Endocrinology and Infertility, Department of Obstetrics and*

524 *Gynecology, University of Texas Health Science Center in San Antonio, TX, USA*

525 Yetunde O Ibrahim

526

527 *Division of Urology, Department of Surgery, University of Utah School of Medicine, Salt Lake*

528 *City, UT, USA*

529 Jim M Hotaling, Kenneth I Aston, Tim G Jenkins

530

531 *Contributions:*

532 TGJ and KIA conceived and designed the study. RHM, CAP, and KB performed all analyses.

533 RHM, CAP, KRB, KIA, and TGJ wrote the manuscript. All authors read and approved the final

534 manuscript.

535 *Corresponding Author:*

Tim G Jenkins, tim_jenkins@byu.edu

Ethics Declarations:

Ethics approval and consent to participate:

All research and sample procurement was conducted in accordance with the regulations and guidelines established by the Institutional Review Boards of University of Utah (IRB_00012049, University of Utah) and Baylor College of Medicine (H-50666, Baylor College of Medicine). Consent was obtained from all participants prior to their donation of tissue samples. All other data sets were publicly available.

Consent for publication:

Not Applicable

Competing interests:

The authors declare the following potential competing financial interests:

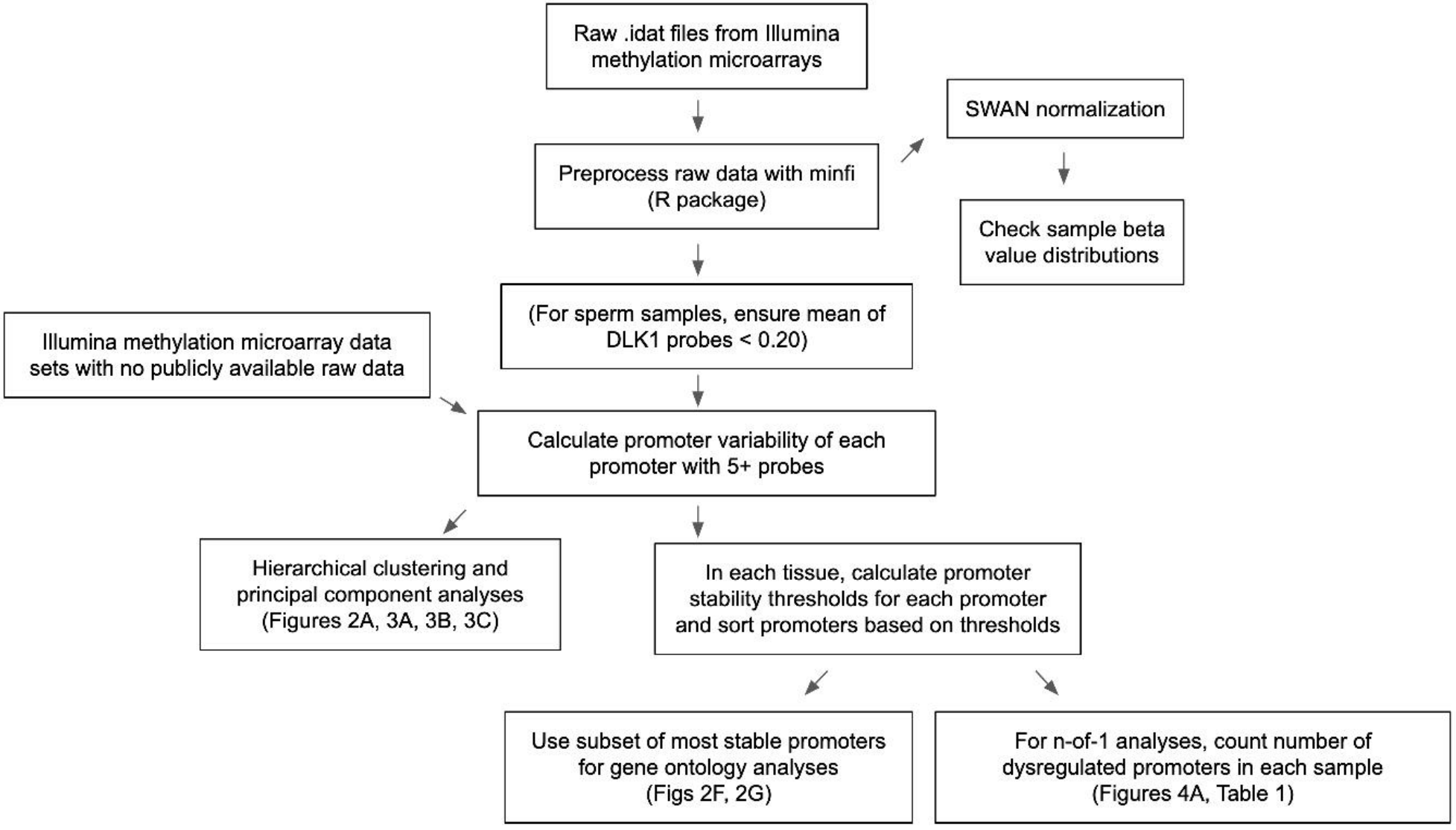
RHM is an employee of Inherent Biosciences. KRB and ACO are co-founders of Inherent Biosciences. TGJ, KIA, LIL, RHM, and JMH have minor stock ownership in Inherent Biosciences. CAP works with Inherent Biosciences via a part-time internship program. A provisional patent on this work was filed prior to manuscript submission. Remaining authors declare no competing interests.

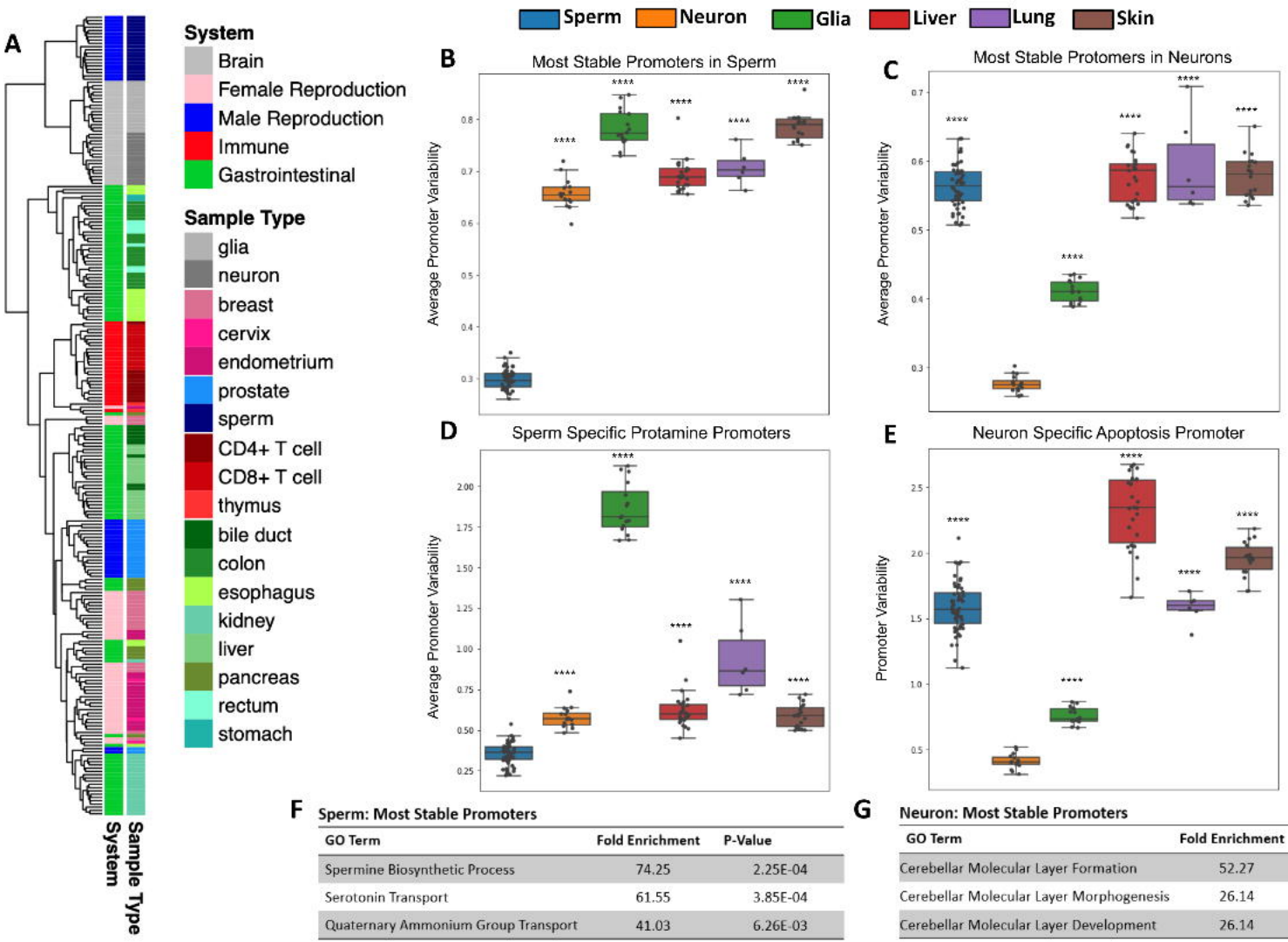
Supplementary Information:

Additional File 1

Supplementary figures and tables

22


Raw .idat files from Illumina methylation microarrays
SWAN normalization
Preprocess raw data with minfi (R package)
Check sample beta value distributions
(For sperm samples, ensure mean of DLK1 probes < 0.20)
Illumina methylation microarray data sets with no publicly available raw data
Calculate promoter variability of each promoter with 5+ probes
Hierarchical clustering and principal component analyses (Figures 2A, 3A, 3B, 3C)
In each tissue, calculate promoter stability thresholds for each promoter and sort promoters based on thresholds
Use subset of most stable promoters for gene ontology analyses (Figs 2F, 2G)
For n-of-1 analyses, count number of dysregulated promoters in each sample (Figures 4A, Table 1)

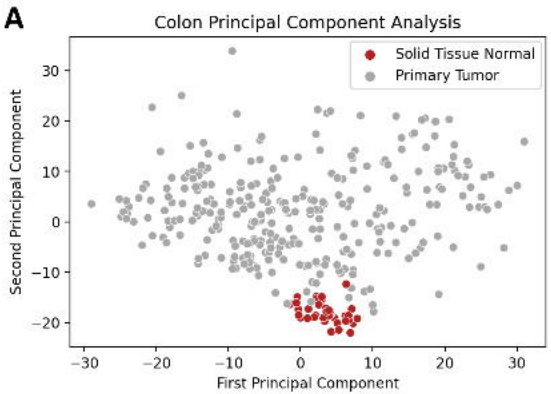


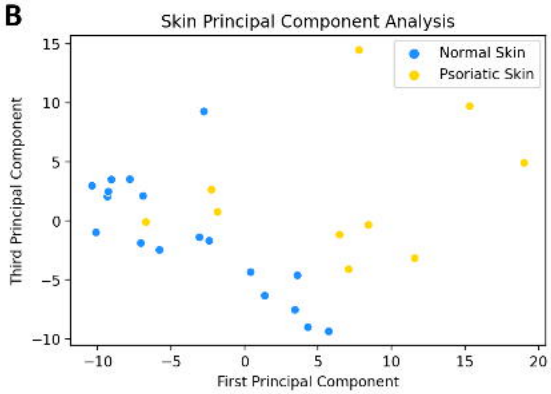
F **Sperm: Most Stable Promoters**

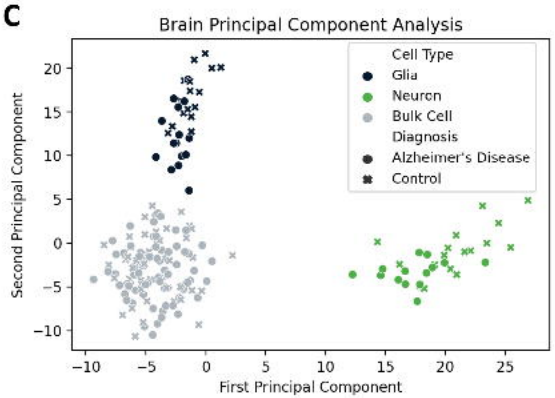
GO Term	Fold Enrichment	P-Value
Spermine Biosynthetic Process	74.25	2.25E-04
Serotonin Transport	61.55	3.85E-04
Quaternary Ammonium Group Transport	41.03	6.26E-03

G **Neuron: Most Stable Promoters**

GO Term	Fold Enrichment	P-Value
Cerebellar Molecular Layer Formation	52.27	3.72E-02
Cerebellar Molecular Layer Morphogenesis	26.14	5.53E-02
Cerebellar Molecular Layer Development	26.14	5.53E-02


A
Colon Principal Component Analysis
Solid Tissue Normal
Primary Tumor
Second Principal Component
First Principal Component


B
Skin Principal Component Analysis
Normal Skin
Psoriatic Skin
Third Principal Component
First Principal Component


C
Brain Principal Component Analysis
Cell Type
Glia
Neuron
Bulk Cell
Diagnosis
Alzheimer's Disease
Control
Second Principal Component
First Principal Component

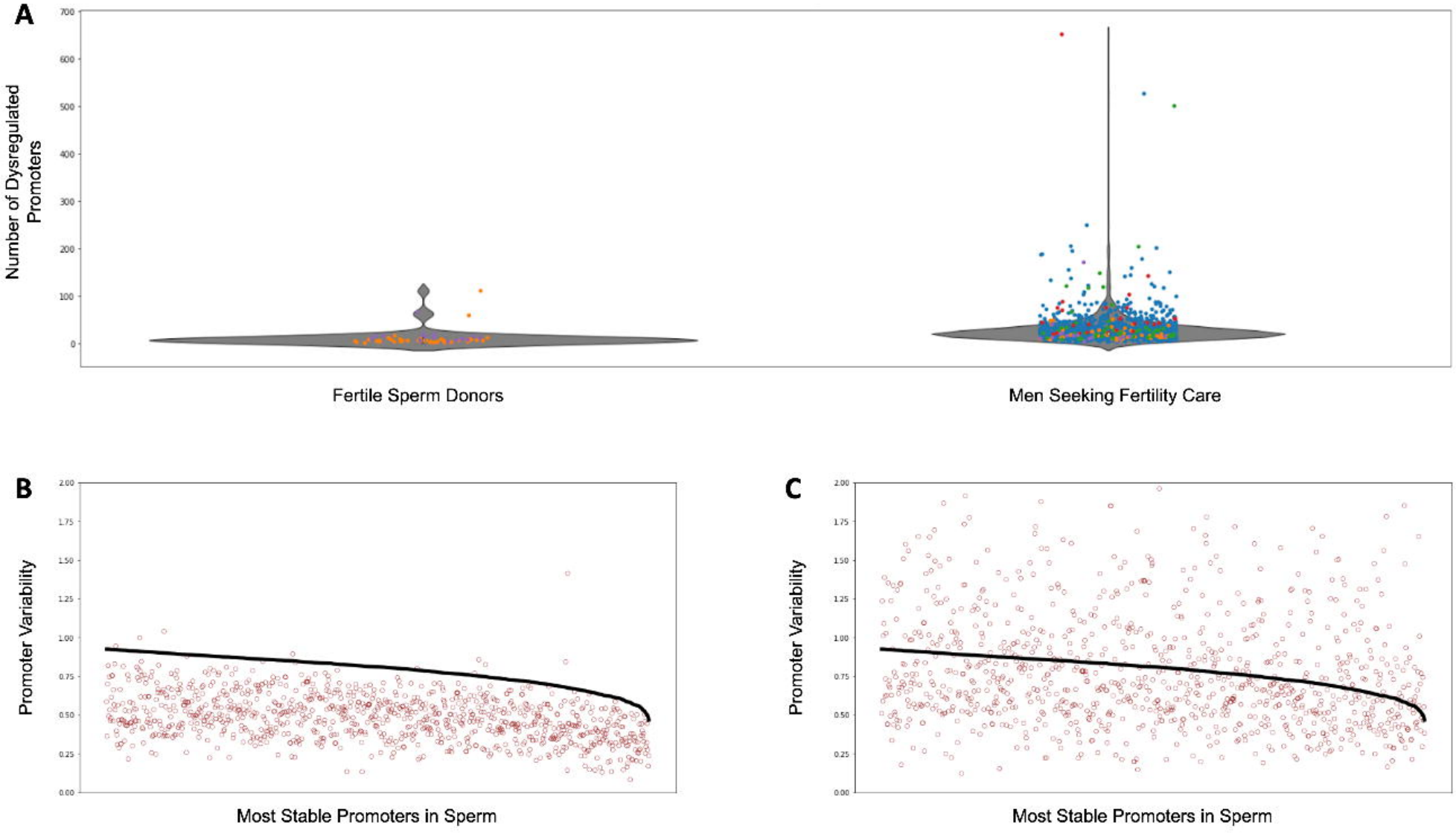

A
Number of Dysregulated Promoters
700
600
500
400
300
200
100
0
Fertile Sperm Donors
Men Seeking Fertility Care
B
Promoter Variability
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Most Stable Promoters in Sperm
C
Promoter Variability
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Most Stable Promoters in Sperm

Table 1A

Patient Cohort	Pregnancy Rate from IUI	Live Birth Rate from IUI
Patients with the least number of dysregulated promoters (bottom 10th percentile, n=54)	48.1%	40.7%
Patients with the most number of dysregulated promoters (top 10th percentile n=50)	26.0%	18.0%
	p=0.019	p=0.011

Table 1B

Patient Cohort	Pregnancy Rate from IVF	Live Birth Rate from IVF
Patients with the least number of dysregulated promoters (bottom 10th percentile, n=36)	75.0%	61.1%
Patients with the most number of dysregulated promoters (top 10th percentile, n=23)	73.9%	60.9%
	p=0.927	p=0.986