Title: Germ line aging and regional epigenetic instability: age prediction using human sperm DNA methylation signatures

Authors: Timothy G Jenkins[1] PhD, Kenneth I Aston[1] PhD, Douglas T Carrell[1,2,3] PhD.

[1]Andrology and IVF Laboratories, Department of Surgery, [2]Department of Obstetrics and Gynecology, [3]Department of Genetics, University of Utah School of Medicine, Salt Lake City, Utah, USA

Corresponding Author: Timothy G Jenkins

675 Arapeen Dr. Suite 201 Salt Lake City, UT, 84108 USA.

tim.jenkins@hsc.utah.edu

801-581-3740.

Abstract:

Background: The relationship between aging and epigenetic profiles has been highlighted in many recent studies and models using somatic cell methylomes to predict age have been successfully constructed. However, gamete aging is quite distinct and as such age prediction using sperm is ineffective with current techniques.

Results: We have produced a model that utilizes human sperm DNA methylation signatures to predict chronological age by utilizing methylation array data from a total of 329 samples. The dataset used for model construction includes infertile patients, sperm donors, and individuals from the general population. Our model is capable of accurately predicting age with an r2 of 0.928 in our test data set. We additionally investigated the repeatability of prediction by processing the same sample on 6 different arrays and found very robust age prediction with an average standard deviation of only 0.877 years. Additionally, we found that smokers have approximately 5% increased age profiles compared to 'never smokers.'

Conclusions: The aging calculator we have built offers the ability to assess "germ line age" by accessing genomic regions affected by age. Our data suggest that this model can predict an individual's chronological age with a high degree of accuracy regardless of fertility status and with a high degree of repeatability. Additionally, our data appear to show age acceleration patterns as a result of smoking suggesting that the aging process in sperm may be impacted by environmental factors though this effect appears to be quite subtle.

Key words: Sperm Epigenetics, aging, DNA methylation, aging calculator.

INTRODUCTION

In the very recent past a great deal of work has been performed in an effort to understand the nature of aging, the mechanisms that drive the process, and the biomarkers that may be predictive of, or affected by age. In this effort, a seminal manuscript was published in 2013 which described the ability to use DNA methylation signatures in somatic tissues to predict an individual's chronological age [1]. In this work, Dr. Horvath demonstrated that the epigenetic mechanisms that reflect the aging process are tightly conserved between individual tissues and across multiple species. This finding was, for many reasons, quite remarkable, not the least of which is the significant contrast in epigenetic profiles between various tissues. Clearly, the aging process is one that affects all tissues in the body and so the similarities in signatures that are predictive of this aging pattern perhaps should have not been too surprising.

Despite the general applicability of this model among various tissues, one tissue in particular did not display similar predictive power as was seen with most. In fact testicular tissue and sperm specifically did not appear to be predictive of age at all with the previously described calculator. In our own unreported trials, human sperm DNA methylation profiles do not offer predictive power to determine one's age using this model. In many ways such a finding may have been expected, as this is not the first finding where the male germ line did not follow typical trends in the aging process. Our lab has previously reported that the nature of alterations to sperm DNA methylation signatures associated with age are opposite of what is typically seen in somatic cells [1-4]. Specifically, we demonstrated that while aging results in a global decrease in

methylation and increased regional methylation in most cells, in sperm the opposite was seen. In addition to this case where the male gamete defies conventional age-associated cellular alterations is telomere length. In fact, while most somatic cells experience marked telomere shortening as a hallmark of aging, sperm do not follow the same trend [5]. Clearly, sperm cells are extraordinarily unique and thus require a unique approach to understand both the nature of the aging process and the potential for some of the biomarkers for aging in the sperm to be predictive of an individual's age as it appears that the aging clock in sperm may be entirely unique.

In our previous publications we have described the general trends of aging on the sperm methylome. In these studies, we have shown that sperm have a very distinct pattern of age-associated alteration [2, 3]. We identified over 140 genomic regions (~1kb in size) that displayed differential methylation with age. Of these, only 8 displayed an increase in methylation, and the remainder showed a marked loss of methylation. Intriguingly, these regions of differential methylation appeared to be enriched at genes known to be associated with bipolar disorder and schizophrenia, both diseases known to have increased incidence in the offspring of older fathers. Indeed the epigenetic patterns of aging in sperm, while distinct from the epigenetic patterns of aging in somatic tissues, are striking and extremely consistent and thus provide an excellent opportunity to utilize in a model designed to identify an individual's age.

The pursuit of generating a model to predict an individual's age using the sperm methylome is not only an interesting question from the perspective of the basic sciences,

but the patterns of sperm aging, and the unique nature of the sperm make the utilization of this cell type ideal for such a predictive model. Using pure cell populations is ideal for any epigenetic analysis, and while the previously constructed models are effective at predicting age even with tissues that are difficult to purify (which is a testament to quality of model and to the strength of the aging signal), the ideal scenario would be to use a more pure cell population. Human sperm offer just such an opportunity. Many protocols are applied to somatic cell removal in sperm epigenetic studies and they have proven quite effective at isolating only the germ cells, thanks in large part to the highly unique and compact nature of the sperm nucleus/head. Further, the magnitude of the aging signal is quite strong in the sperm (thought to be in part due to the highly proliferative nature of the sperm cells themselves) and as a result, the patterns of aging offer an excellent opportunity for predictive power. In this study, we set out to capitalize on these advantages to build a model that can predict an individual's age using methylation signatures in their germ line. We have also designed experiments to inform us of the actual impact of this prediction and if alterations in an individual's predicted age may be the result of environmental exposures or lifestyles (smoking, obesity, etc.).

RESULTS

Model construction and training: In the current study we assessed sperm DNA methylation array data (Illumina 450K array) from 3 distinct previously performed studies [2, 6, 7]. From these data sets, we were able to acquire a total of 329 samples that were used to generate the predictive model outlined herein. Individuals with many different fertility phenotypes provided the samples used in this study. Specifically, our

training data set includes samples from sperm donors, known fertile individuals, infertility patients (including those seeking intrauterine insemination or even in vitro fertilization treatment at our facility), and individuals from the general population. Further, our data set includes those that have very different lifestyles and environmental exposures (as an example, both heavy smokers and never smokers are represented in our data set).

We utilized the glmnet package in R to facilitate training and development of our linear regression age prediction model [8]. For training of our model, we limited the training dataset to only 147 regions that we have previously been identified to be strongly associated with the aging process to ensure a more clear interpretability to the results of the model [2]. We trained multiple models to identify the best possible outcomes. First, we trained on all of the beta-values for each CpG located in our regions of interest ("CpG level" training). Second, we generated a mean of beta-values for each region, which included the CpGs within each region respectively. Ultimately, this approach yielded a list of mean beta-values for each region ("regional level" training), and the model was trained only on these averages.

In each of the above-described scenarios, we employed a 10-fold cross validation strategy. This was performed 10 times on unique subgroups of the entire data set (Figure 1A-F). The results from these ten validations were compared between the CpG level training and the regional level training. To compare the accuracy and predictive power of these models we performed linear regression for each (actual age vs. predicted age) and

generated r^2 values. These r^2 values were compared via simple two-tailed t-test to determine if any significant differences exist between the two approaches to model construction (CpG level construction vs. regional level construction). These tests revealed that there was a moderately significant decrease in predictive power in the regional model when considering only the training data sets ($p=0.0428$). However, there was no significant difference seen between the test sets ($p=0.3439$). In fact, while significant, the CpG level model appeared to be more prone to extremes in significantly lower predicative power in individual test sets when compared to the regional level models (Figure 1G). In an effort to make the model as simple as possible and in light of these findings, we committed to use the regional level model moving forward.

We additionally assessed the weighting of the features (regions) used in the models constructed during cross validation. We found a great deal of variation in the features selected across the regions screened, though a certain percentage of the regions were heavily weighted and used in 80% or more of the models built during cross validation (a total of 51 features/regions met this criterion). In an effort to avoid over-fitting we compared cross validation (10-fold strategy) only in these 51 regions ("optimized regions") to all of the regions previously screened. We found that both the training and test groups were not statistically different between the optimized regional list and the full regional list (Figure 1H). We therefore selected the best performing model that was trained only on 51 regions of the genome (Table 1). With this model we are able to generate a prediction of all 329 samples in our data set with an r^2 of approximately 0.89

and an average accuracy in prediction of 93.7% (calculated as the average of |(prediction/actual age)-1|).

Technical validation / replicate performance: Because array-to-array variability can be a concern, we tested our model in a completely independent cohort of samples (different samples from different batches of arrays). We were able identify 10 sperm samples each with 6 technical replicates that were each run on the 450K array (not those used in our cross validation / model training) to determine the precision and consistency the model has in predicting an individuals age. The model performed well with the average standard deviation in age prediction being only 0.877 years and a regression analysis (predicted vs. actual age) revealing an r^2 of 0.7319 and a p=0.0016 (Figure 2).

Association with regional inter-individual epigenetic instability: To determine the drivers of the age-associated change in these regions, we examined the level of variability/instability of methylation signals between the individual's screened in our test/training set. We assessed the methylation variability between individuals at each of these sites by methods we have previously described [7]. Using these techniques we binned individuals by year of age and then generated measures of variation within each bin (to avoid the increased levels of variation that would be present simply due to the change with age we have identified). We then tested the difference between variability at the age-associated regions within each bin against the background of the entire array. We found a significantly increased variability on average at age-associated regions compared to background at nearly every age (Figure 3). We also found that highly variable

methylation signals at single CpGs is relatively common throughout the array and that the higher average levels of variability within age-associated regions were not defined by CpGs with variability levels elevated above what is seen elsewhere. Instead, the driver of increased averages seen in our analysis of age-associated regions was due to an increase in the frequency of these high variable CpGs, not in the magnitude of their variation (Figure 3A).

The impact of smoking on age prediction: To test the potential diagnostic/clinical utility of our model we have more closely assessed the data in our original cross validation dataset. Specifically we have analyzed our smoking dataset, which includes sperm methylation data from 78 never smokers and 78 smokers. Similar aged men are represented in each group. We additionally isolated a portion of the smoking group who were considered 'long term smokers' for analysis (>10 years consuming cigarettes). We found an approximately 1.5% increased in predicted age compared to chronological age in all smokers and 2.5% increase in long term smokers. However this difference failed to reach statistical significance. Interestingly, this same pattern was observed (though significantly higher in magnitude) when screening only individuals who were less than 35 years old at the time of collection (Figure 4). In these samples we saw a 3% increase in predicted age compared to chronological age in the smoker group and a nearly 6% increase in predicted age in the long-term smokers (p=0.0196).

DISCUSSION

We have developed a sperm age calculator that has the capacity to identify an individual's chronological age based only on their sperm DNA methylation signatures. While previous studies have very successfully generated an aging calculator for somatic cells, these calculators fail to work effectively with germ line epigenetic signatures. Herein we have described the development of a linear model that has the ability to accurately predict ages with these signatures. Specifically our model is based on average methylation signatures at 51 genomic loci known to be altered as men age [2].

In the process of model construction, we evaluated multiple potential methods by which we could train our model. One important consideration was the nature of the population with which the model was trained. While there is a balance in selecting your population (broad applicability vs targeted population) we decided to utilize a population with diverse fertility phenotypes and exposers to ensure that it could perform well with many different phenotypes. As such, we included, smokers and non-smokers, Individuals of known fertility, those currently being treated for infertility, and men from our general population. By doing so we sought to ensure that our model was as broadly applicable as possible.

We also spent a great deal of effort to ensure that the model was as simple as possible. While training on all data from the entire array may have provided additional power in prediction, it also would likely make the model very difficult to interpret. Instead, we focused only on the regions that we knew were independently predictive of age (based on previous data) and refined the model by only assessing these regions. In fact, we found

that even in our simplified model there was some degree of over-fitting that was occurring, and we were able to further simplify form our initial 147 regions down to 51 regions with just as high of predictive power. This effort resulted in a quite robust model (~94% accuracy and an r2 of ~0.89). If our final model had a great deal of room for improvement, there would be a larger need for revisiting our approach and potentially increasing our training feature set to include more, or all, of the array. However, since we have such a robust and interpretable model as it stands, pursuing a different course was not warranted.

Our data indicate that the model constructed herein is also technically robust. We were able to assess previous data from our lab in which 10 individuals had six technical replicates on 450k methylation arrays [9]. This replicate data enabled to assess the power of the model in two distinct ways. First, we were able to assess the predictive power of the model on a completely independent cohort (each of these samples were performed at a different time and on different arrays than what the model was trained upon). Second, we were able to show that the model is able to generate consistent predictions for individuals between technical replicates. Of additional interest is the fact that the samples used in these technical replicates originated from a study that tested the impact of extreme and prolonged temperature exposures on sperm DNA methylation patterns. Thus a portion of the replicates screened were exposed to various magnitudes less than ideal conditions, adding further validity to the strength of the aging signal in the sperm methylome and ultimately to this model. This stability between various batches and samples is important in a model that will have broad applicability.

Our study also indicates that there is significant increases in inter-individual epigenetic instability / variability at sites known to be affected by age. To perform this experiment we had to be careful to only observe variability measures within a fixed age (as the change in methylation over time at age-affected regions would clearly result in increased variability if observed across multiple ages as we know these sites will change). However, when increased variability is observed at these sites when age is held constant (with only a one year window being considered) it suggests that there is a real biological instability at these sites between individuals. Multiple potential explanations exist for this finding, the simplest of which is that even within a single year's time, the miniscule change that occurs during that period results in elevated variability. This seems unlikely based on our data because we commonly found similar and even higher levels of CpG methylation variability across the genome in non-age-affected regions. While not increased in magnitude at the age-affected sites, there was an increase in the frequency of CpGs with elevated variability, which was the driver for the increased average variability across the age-affected regions. Another potential explanation for this increase in variability is that fathers conceiving offspring at different ages may pass on some of these marks to the offspring. At face value, this seems a bit far-fetched due to the massive reprograming events, which take place in the early embryo and in the primordial germ cells. However, there are data available that suggest that methylation marks in many sub-telomeric regions escape reprograming events and can be potentially be passed on to the offspring [10-14]. Intriguingly, our original sperm aging study showed that the majority of age-affected regions were located in these sub-telomeic regions as well [2]. Such a

transmission of age-affects would be remarkable, but may offer a real potential explanation for at least a portion of the variation seen at these regions. Regardless of the mechanism driving the increased variability, these patterns are intriguing and warrant further study.

Our data also suggest that there may be some utility for such a model in a clinical setting. Specifically, we were able to identify an age-affect of smoking in our cohort of patients. We found that individuals who smoke appeared to have acceleration in the pattern of aging and thus the individual's "germ line age" was in some cases significantly higher than their chronological age. This represents one example of many different analyses that could be performed and we may find that different levels/types of infertility, obesity, or other environmental exposures may cause acceleration in the aging pattern seen in sperm. One of the biggest questions that remain if such a finding is real is the potential impact of this age acceleration. Such a pattern could potentially result in increased risk to offspring health as epidemiological data clearly shows increased incidence of neuropsychiatric disease in the offspring of older fathers [15-20]. Such an increase in risk may not mean that the altered methylation pattern itself causes these offspring abnormalities, but instead the methylation signatures of age are simply a good indicator of the overall state or age of the sperm. Likely of more immediate interest to clinicians is the fact that advanced paternal age is associated with a loss of fecundity and fertility. Specifically, it has been shown that men older than 45 years take ~5 times as long to achieve a pregnancy as men less than 25 years (when controlling of female age) [21]. A similar decrease in fecundity was identified in a large population study in 2000 which showed that (after adjusting for

maternal age) men > 35 years had a 50% lower chance of achieving a pregnancy within 12 months of attempting conception [22]. Other studies have also shown decreased fertilizing potential in both IUI and IVF [23, 24]. While the magnitude of this effect remains controversial [25, 26], it is clear that advanced paternal age does play an important role in a couple's fertility status and can clearly result in, at a minimum, a significantly increased time to pregnancy. For many couples, such potential barriers to achieving a pregnancy are essential to understand and discuss with their care providers. While none of these associations have been proven in this specific work, the potential clinical utility of the calculator is intriguing and warrants further investigation both in the individual's health/fertility as well as in the prediction of sired offspring phenotype.

The data described herein are quite promising, though some limitations are clear. Foremost among them are our knowledge of downstream impacts as described above. This will require a great degree of effort to determine what the nature of these effects truly are and if risks can be modified in any way by various treatments. Further, while the current model is very effective at predicting an individuals age and is quite robust technically, the alterations we are observing to predict age are subtle and thus small inefficiencies can result in an inability to detect meaningful changes. Despite this, because of the approach we have taken in designing a model based only on limited numbers of regions there is a potential to modify this model for use with a different platform, namely targeted sequencing. With a targeted sequencing approach, we may be able to improve an already robust predictive model by multiplex sequencing with extreme depth at only the 51 sites of interest. This could provide an even more economical and

reliable predictive model. Taken together, the data that we have shown here are intriguing and warrant a great deal of further investigation and we also have the potential to improve predictive power with future iterations.

METHODS

Samples, study design, data availability: In the current study we assessed sperm DNA methylation array data from 3 distinct previously performed studies [2, 6, 7]. All of the studies have been previously performed in our laboratory. We included only the samples for which ages for the individuals tested were available. From these data sets, we were able to acquire a total of 329 samples that were used to generate the predictive model outlined herein. Each sample was run on the Illumina 450K methylation array. In each case we used SWAN normalization to generate beta-values (values between 0 and 1 that represent the fraction of a given CpG that is methylated) that were used During the early processing of the sperm samples, great care was taken to ensure that there no somatic cell contamination was present that could potentially influence the results of our studies. To prove that this has effectively taken place we assessed the methylation signatures at a number of sites throughout the genome, each of which are highly differentially methylated between sperm and somatic tissues. In Figure 5, we show the differential methylation at one representative genomic locus, DLK1, to illustrate the absence of contaminating signals in the samples used in our study. While a great degree of variability exists between the methylation in these samples there exists very little, if any somatic methylation signals. This pure population is key to ensuring a robust model most targeted at detecting the actual variable of interest, age.

Samples used: We selected the groups to be used for training the predictive model for a few distinct reasons. Individuals with many different fertility phenotypes provided the samples used in this study. Specifically, our training data set includes samples from sperm donors, known fertile individuals, infertility patients (including those seeking intrauterine insemination or even in vitro fertilization treatment at our facility), and individuals from the general population. Further, our data set includes those that have very different lifestyles and environmental exposures (as an example, both heavy smokers and never smokers are represented in our data set).

Model Training: We utilized the glmnet package in R to facilitate training and development of our linear regression age prediction model [8]. For training of our model, we limited the training dataset to only 147 regions that we have previously identified to be strongly associated with the aging process to ensure the broad interpretability to the results of the model [2]. We trained multiple models to identify the best possible outcomes. First, we trained on all of the beta-values for each CpG located in our regions of interest ("CpG level" training). Second, we generated a mean of beta-values for each region which included the CpGs within each region respectively. Ultimately, this approach yielded a list of mean beta-values for each region ("regional level" training), and the model was trained only on these averages.

In each of the above-described scenarios, we employed a 10-fold cross validation strategy to repeatedly test trainings on 90% of our samples and hold out 10% for a test set. This

was performed 10 times on unique subgroups of the entire data set. The results from these ten validations were compared between the CpG level training and the regional level training. To compare the accuracy and predictive power of these models we performed linear regression for each (actual age vs. predicted age) and generated r^2 values. These r^2 values were compared via simple two-tailed t-test to determine if any significant difference exists between the two approaches to model construction (CpG level construction vs. regional level construction).

Technical validation / replicate performance: We tested our model in a completely independent cohort of samples [9]. We were able identify 10 sperm samples each with six technical replicates that were that were each run on the 450K array (not those used in our cross validation / model training) to determine how precise and consistent the model is at predicting an individuals age. These samples were run with the final predictive model to and a linear regression analysis of predicted vs. actual age was performed using R.

Association with regional inter-individual epigenetic instability: To determine the drivers of the age-associated change in these regions, we examined the level of variability/instability of methylation signals between the individual's screened in our test/training set by a method previously described in a recent publication from our lab [7]. We assessed the methylation variability between individuals at each of these sites by methods we have previously described [7]. In brief, the variability analysis begins with logit transformation of beta values to ensure homoscedasticity followed by a center scaling (using the 'scale' function in R), which generates a distance from the average for

each CpGs. The absolute center scaled value is then used to determine the absolute distance from the mean and is averaged across the age effected regions and background CpGs (all other CpGs on the array). We then compare the center scaled values in these two groups (age affected and background) to determine if there is elevation in variability or distance from the mean. In this specific ananlysis we binned individuals by year of age and then performed center scaling within each bin (to avoid the increased levels of variation that would be present simply due to the change with age we have identified). We then tested the difference between variability at the age-associated regions within each bin against the background of the entire array.

The impact of smoking on age prediction: To test the potential diagnostic/clinical utility of our model we have more closely assessed the data in our original cross validation dataset. Specifically we have analyzed our smoking dataset, which includes sperm methylation data from 78 never smokers and 78 smokers. Similar aged men are represented in each group. We additionally isolated a portion of the smoking group who were considered ‘long term smokers’ for analysis (>10 years consuming cigarettes). In this analysis we compared accuracy of the age prediction of each group to determine if there is a significant increase in the age prediction compared to chronological age in individuals who smoke. To do this we identified we compared predicted age vs. actual age by the equation % difference = (predicted age/actual age)-1. We then compared the % difference values for each group via two-tailed t-test to determine if there were significant differences in age acceleration between the two groups.

ACKNOWLEDGMENTS:

We are grateful for the kind assistance of Andrew Smith PhD from USC expertise in machine learning techniques and model training. Similarly we recognize the efforts of Chris Conley from the Huntsman Cancer Institute for his technical assistance.

FUNDING:

A portion of the data used in this manuscript originated from work performed in our lab, which was funded by the NIH (RO1HD082062).

REFERENCES:

1. Horvath S: **DNA methylation age of human tissues and cell types.** *Genome Biol* 2013, **14:**R115.
2. Jenkins TG, Aston KI, Pflueger C, Cairns BR, Carrell DT: **Age-associated sperm DNA methylation alterations: possible implications in offspring disease susceptibility.** *PLoS Genet* 2014, **10:**e1004458.
3. Jenkins TG, Aston KI, Cairns BR, Carrell DT: **Paternal aging and associated intraindividual alterations of global sperm 5-methylcytosine and 5-hydroxymethylcytosine levels.** *Fertil Steril* 2013, **100:**945-951.
4. Richardson B: **Impact of aging on DNA methylation.** *Ageing Res Rev* 2003, **2:**245-261.
5. Allsopp RC, Vaziri H, Patterson C, Goldstein S, Younglai EV, Futcher AB, Greider CW, Harley CB: **Telomere length predicts replicative capacity of human fibroblasts.** *Proc Natl Acad Sci U S A* 1992, **89:**10114-10118.
6. Aston KI, Uren PJ, Jenkins TG, Horsager A, Cairns BR, Smith AD, Carrell DT: **Aberrant sperm DNA methylation predicts male fertility status and embryo quality.** *Fertil Steril* 2015, **104:**1388-1397 e1381-1385.
7. Jenkins TG, James ER, Alonso DF, Hoidal JR, Murphy PJ, Hotaling JM, Cairns BR, Carrell DT, Aston KI: **Cigarette smoking significantly alters sperm DNA methylation patterns.** *Andrology* 2017.
8. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *J Stat Softw* 2010, **33:**1-22.
9. Jenkins TG SM, James E, Aston KI, Carrell DT: **Thermo stability of DNA methylation marks in human sperm.** *J Genet and Genome Res* 2017, **3**.

10. Guibert S, Forne T, Weber M: **Global profiling of DNA methylation erasure in mouse primordial germ cells.** *Genome Res* 2012, **22:**633-641.
11. Hajkova P, Erhardt S, Lane N, Haaf T, El-Maarri O, Reik W, Walter J, Surani MA: **Epigenetic reprogramming in mouse primordial germ cells.** *Mech Dev* 2002, **117:**15-23.
12. Franklin TB, Russig H, Weiss IC, Graff J, Linder N, Michalon A, Vizi S, Mansuy IM: **Epigenetic transmission of the impact of early stress across generations.** *Biol Psychiatry* 2010, **68:**408-415.
13. Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, Jacobsen SE, Reik W: **Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency.** *Nature* 2010, **463:**1101-1105.
14. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR: **Transgenerational epigenetic instability is a source of novel methylation variants.** *Science* 2011, **334:**369-373.
15. Frans EM, Sandin S, Reichenberg A, Langstrom N, Lichtenstein P, McGrath JJ, Hultman CM: **Autism risk across generations: a population-based study of advancing grandpaternal and paternal age.** *JAMA Psychiatry* 2013, **70:**516-521.
16. Idring S, Magnusson C, Lundberg M, Ek M, Rai D, Svensson AC, Dalman C, Karlsson H, Lee BK: **Parental age and the risk of autism spectrum disorders: findings from a Swedish population-based cohort.** *Int J Epidemiol* 2014, **43:**107-115.
17. Miller B, Messias E, Miettunen J, Alaraisanen A, Jarvelin MR, Koponen H, Rasanen P, Isohanni M, Kirkpatrick B: **Meta-analysis of paternal age and schizophrenia risk in male versus female offspring.** *Schizophr Bull* 2011, **37:**1039-1047.
18. Naserbakht M, Ahmadkhaniha HR, Mokri B, Smith CL: **Advanced paternal age is a risk factor for schizophrenia in Iranians.** *Ann Gen Psychiatry* 2011, **10:**15.
19. Dalman C: **Advanced paternal age increases risk of bipolar disorder in offspring.** *Evid Based Ment Health* 2009, **12:**59.
20. Kuratomi G, Iwamoto K, Bundo M, Kusumi I, Kato N, Iwata N, Ozaki N, Kato T: **Aberrant DNA methylation associated with bipolar disorder identified from discordant monozygotic twins.** *Mol Psychiatry* 2008, **13:**429-441.
21. Hassan MA, Killick SR: **Effect of male age on fertility: evidence for the decline in male fertility with increasing age.** *Fertil Steril* 2003, **79 Suppl 3:**1520-1527.
22. Ford WC, North K, Taylor H, Farrow A, Hull MG, Golding J: **Increasing paternal age is associated with delayed conception in a large population of fertile couples: evidence for declining fecundity in older men. The ALSPAC Study Team (Avon Longitudinal Study of Pregnancy and Childhood).** *Hum Reprod* 2000, **15:**1703-1708.
23. Mathieu C, Ecochard R, Bied V, Lornage J, Czyba JC: **Cumulative conception rate following intrauterine artificial insemination with husband's spermatozoa: influence of husband's age.** *Hum Reprod* 1995, **10:**1090-1097.
24. Dain L, Auslander R, Dirnfeld M: **The effect of paternal age on assisted reproduction outcome.** *Fertil Steril* 2011, **95:**1-8.

25. Niederberger C: **Re: Male Biological Clock: A Critical Analysis of Advanced Paternal Age.** *J Urol* 2016, **195:**717.
26. Ramasamy R, Chiba K, Butler P, Lamb DJ: **Male biological clock: a critical analysis of advanced paternal age.** *Fertil Steril* 2015, **103:**1402-1406.

Figure Legends:

Figure 1: (A-F) Scatterplots depicting the relationship between predicted and chronological age in 6 represented models from our cross validation testing. (G) Box and whisker plots of the R2 values from each cross validation (10) for both training and test datasets between the CpGs level data and the regionalized data. (H) Box and whisker plots of the R2 values from each cross validation (10) for both the full regional data set (147 regions) and the optimized regional data set (51 regions) with both training and test data displayed.

Figure 2: (A) scatterplot depicting the age prediction in a completely independent cohort of samples. (B) Boxplots demonstrating the variation in age prediction from ten individuals with six biological replicates that were run in a completely independent cohort.

Figure 3: Figure demonstrating the assessment of epigenetic instability at age-affected regions of the genome. (A) Dot plot depicting the level of methylome instability at CpGs

based on their distance away from the center of age-affected regions with a heat map displaying the fraction of sites that were higher than 1 standard deviation above the average instability value (data were binned in 500bp bins based on distance up or downstream from the center of age affected regions). (B) Bar plot depicting average methylome instability between all CpGs on the array and those within age-affected regions. The data was binned and assessed for instability within a single year. (C) Box and whisker plot depicting the difference in average methylome instability between the entire array and age-affected regions. This difference was statistically significant based on two-tailed t-test ($p<0.00001$).

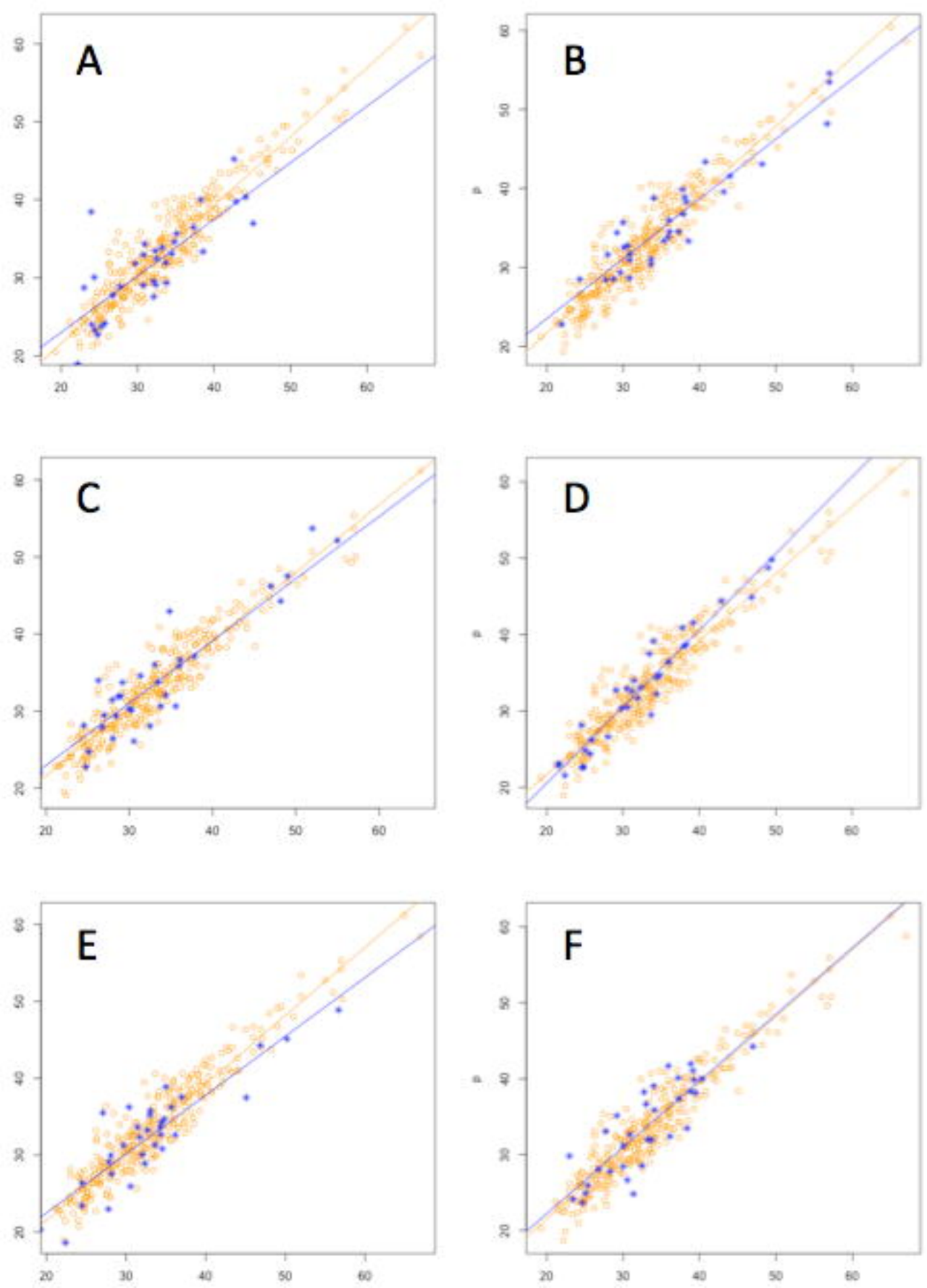
Figure 4: Density plot shows the accuracy of age prediction in never smokers, smokers, and heavy smokers among individuals below 35 years of age. Similar patterns exist in the entire cohort but are the most profound in this age group.

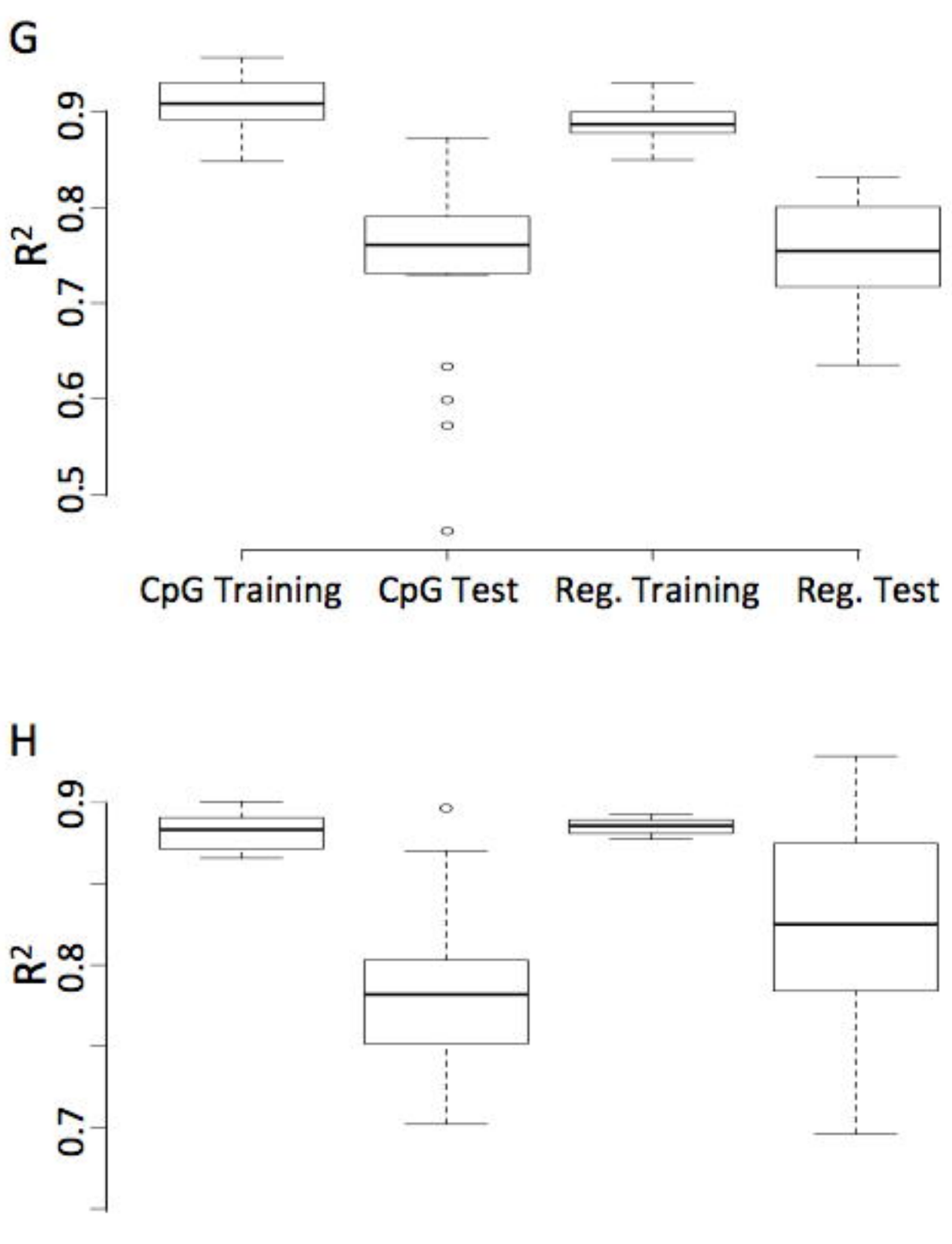
Figure 5: Heatmap of the DLK1 locus, which is highly differentially methylated between sperm and somatic cells is used to confirm the absence of contaminating signals in our data set. 4 blood samples are listed at the far left of the heatmap and the remainder of the samples used in our study follow.

Table 1:

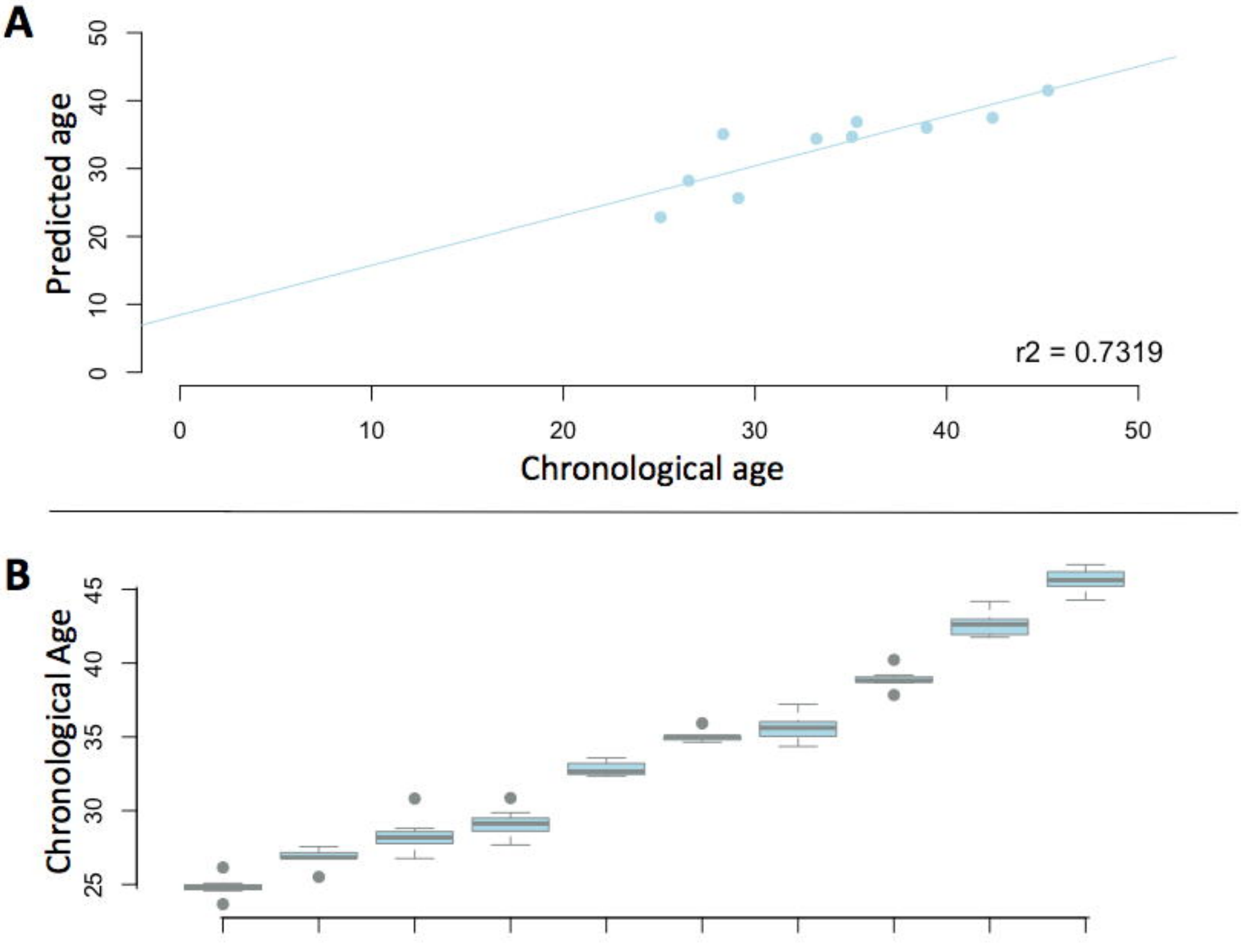
Name	CHR	Start	Stop
ADAMTS8	chr11	130299298	130299948
ARC	chr8	143694010	143694548
ARGHGEF10	chr8	1877888	1878324
BCL11A	chr2	60680616	60680762
C1ORF122	chr1	38272200	38273057
C7ORF50	chr7	1083209	1084163
CCDC144NL	chr17	20798895	20799770
CLIC1	chr6	31698492	31699299
DMPK	chr19	46282571	46283081
FAM86C1	chr11	71498202	71499118
FAM86JP	chr3	125634060	125634453
FOXK1	chr7	4722778	4723928
FSCN	chr7	5635134	5635954
GAPDH	chr12	6641602	6642355
GET4	chr7	914964	915832
GNB2	chr7	100274361	100275305
GPANK1	chr6	31630819	31632542
GPR45	chr2	105857809	105859084
KCNQ1	chr11	2554562	2555577
LDLRAD4	chr18	13611370	13611825
LMO3	chr12	16760040	16761003
LOC100133461	chr4	3680721	3681760
MIR22HG	chr17	1617363	1618296
MTMR8	chrX	63614857	63615496
N10	chr1	28423399	28424202
N12	chr5	3593413	3594276
N22	chr19	4579481	4580471
N23	chr14	106004434	106004608
N24	chr6	170449417	170450804
N27	chr6	30432200	30433944
N30	chr15	27959473	27960032
N8	chr11	69260136	69261045
N9	chr7	35300077	35301070
NCOR2	chr12	124990897	124991140
NONE	chr10	17347047	17347392
NSG1	chr4	4386726	4387698
PAX2	chr10	102509693	102510569
PITX1	chr5	134365728	134366535

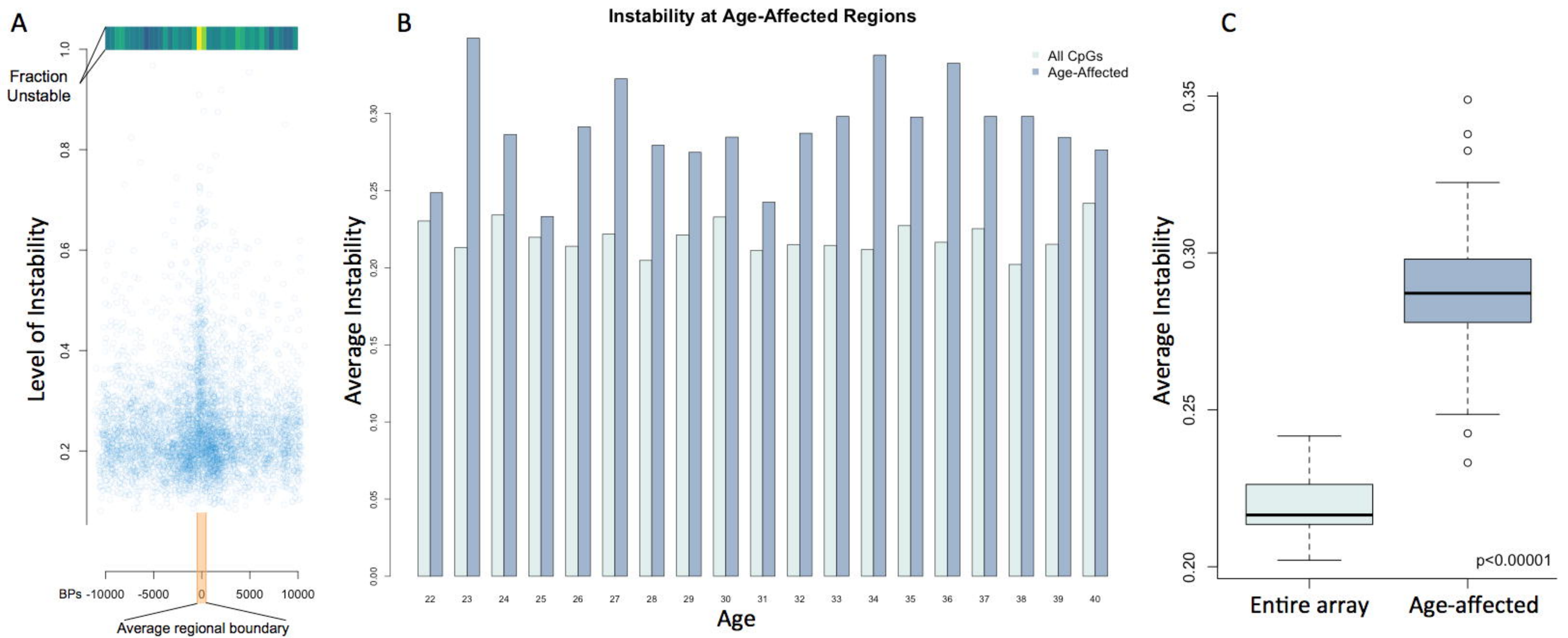
PRSS22	chr16	2908157	2908935
PTPRN2.3	chr7	157523356	157524159
PTPRN2.4	chr7	158109339	158110153
PURA	chr5	139492535	139493491
PYY2	chr17	26553567	26554908
SECTM1	chr17	80278592	80280331
SEMA6B	chr19	4555999	4556983
SEZ6	chr17	27330794	27332647
SLC22A18AS	chr11	2909690	2909716
SOHLH1	chr9	138590204	138590996
THBS3	chr1	155176868	155177784
TNXB	chr6	32064146	32065891

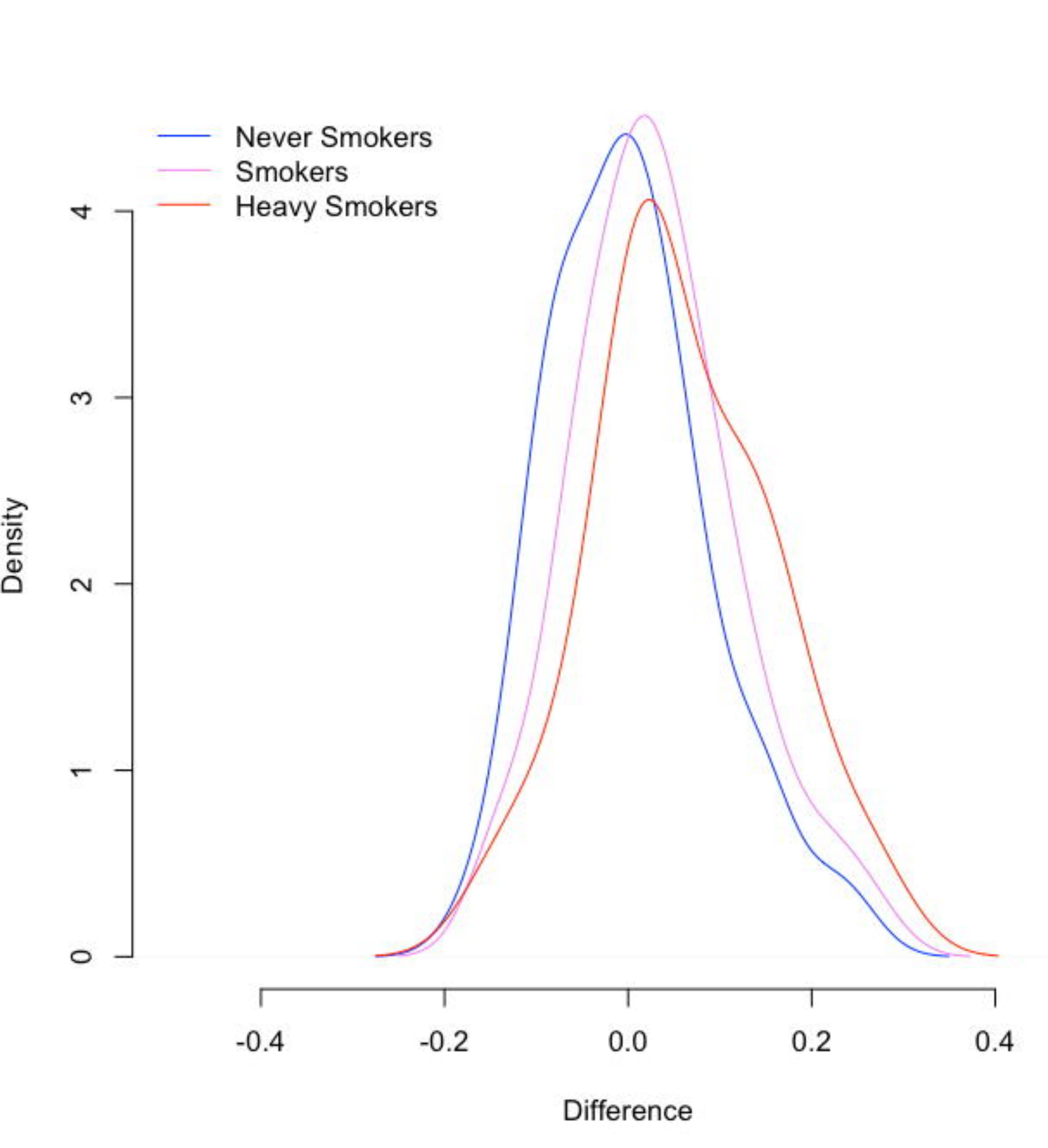

A
B
C
D
E
F

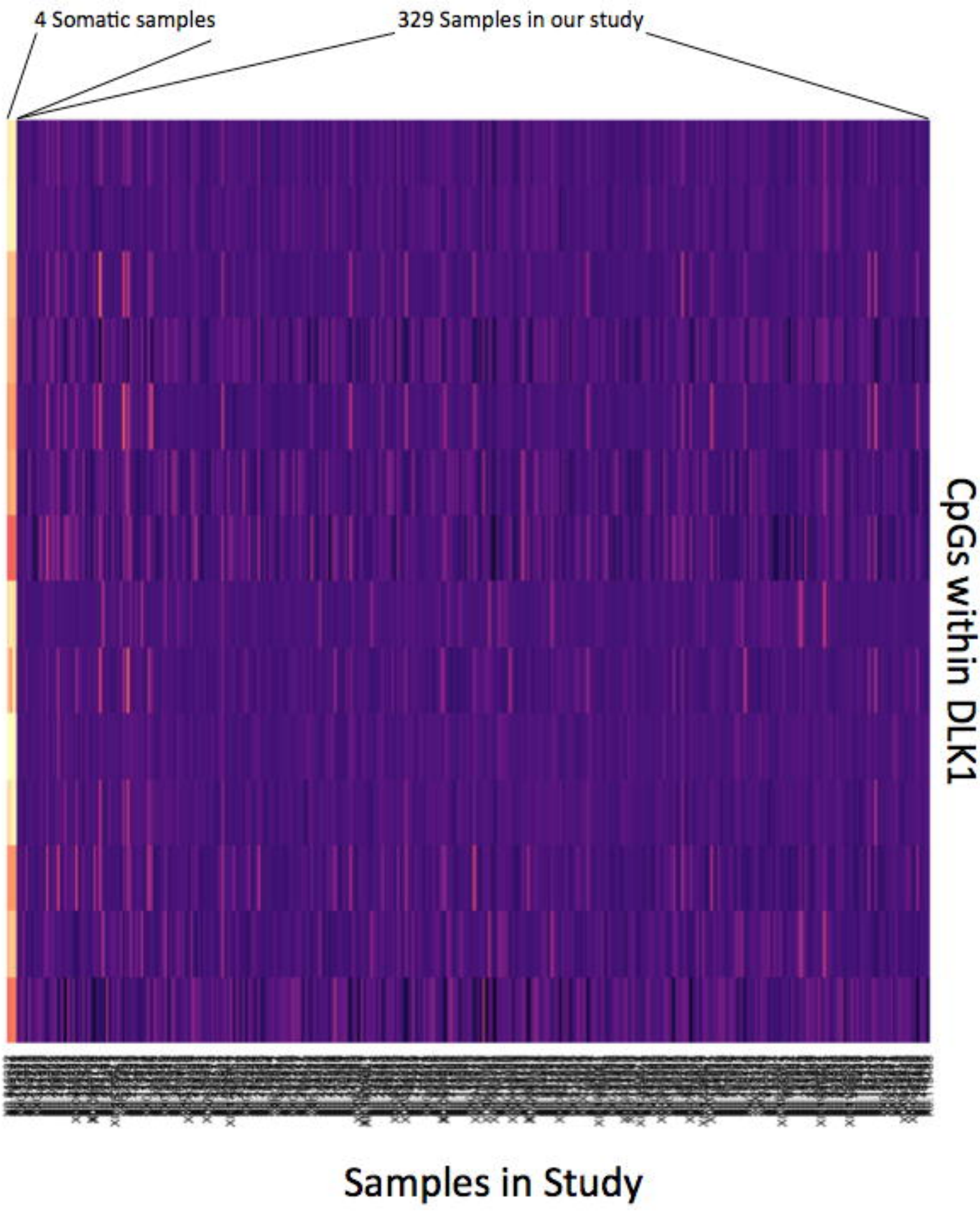

G
R2
0.5
0.6
0.7
0.8
0.9
CpG Training
CpG Test
Reg. Training
Reg. Test

H
R2
0.7
0.8
0.9
Full Training
Full Test
Opt. Training
Opt. Test


A
Predicted age
0
10
20
30
40
50
r2 = 0.7319
0
10
20
30
40
50
Chronological age
B
Chronological Age
25
30
35
40
45


A
Fraction Unstable
Level of Instability
1.0
0.8
0.6
0.4
0.2
BPs -10000
-5000
0
5000
10000
Average regional boundary
B
Instability at Age-Affected Regions
All CpGs
Age-Affected
Average Instability
0.00
0.05
0.10
0.15
0.20
0.25
0.30
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
Age
C
Average Instability
0.20
0.25
0.30
0.35
p<0.00001
Entire array
Age-affected


Never Smokers
Smokers
Heavy Smokers
Density
0
1
2
3
4
-0.4
-0.2
0.0
0.2
0.4
Difference


4 Somatic samples
329 Samples in our study
CpGs within DLK1
Samples in Study