Landes Bioscience

www.landesbioscience.com

Supplemental Material to:

Robert Lowe, Carolina Gemma, Huriya Beyan, Mohammed I. Hawa, Alex Bazeos, R. David Leslie, Alexandre Montpetit, Vardhman K. Rakyan and Sreeram V. Ramagopalan

Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies

Epigenetics 2013; 8(4)
http://dx.doi.org/10.4161/epi.24362

http://www.landesbioscience.com/journals/epigenetics/article/24362/

Table S1: Sample statistics for buccal BS-Seq data

BS-Seq mapping

All BS-Seq samples were initially QC by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and were then mapped using BiFast (https://bitbucket.org/xboxrob/bifast) with parameters –n 1 and –l 50. For blood and buccal data a large percentage of the reverse end of the paired end reads mapped poorly and hence those that did not map were re-mapped using the forward end in single end mode. Paired end reads were filtered for clonality by assigning reads as clonal if the forward and reverse read mapped to exactly the same location. A read from all the clonal reads at each location was taken using random sampling providing it passed a minimum mapping quality of 15. For single end reads, a clonal read was defined by those reads that mapped to the same genomic location. Bisulphite conversion rates were calculated by measuring the conversion of cytosines in a non-CpG context (Buccal 98.1% Blood 99.1%; **Table S2**)

Sample ID	sex	ethnicity	age (years) at buccal cell extraction
1	Male	caucasian	51
2	Male	caucasian	24
3	Male	caucasian	18
4	Male	caucasian	24
5	Male	caucasian	20
6	Male	caucasian	51
7	Female	caucasian	40
8	Male	caucasian	11
9	Male	caucasian	72
10	Female	caucasian	29
11	Female	caucasian	40
12	Male	caucasian	42
13	Male	caucasian	40
14	Male	caucasian	72

Table S2: Pooled statistics for blood and buccal BS-Seq data.

	Number of Sequences	Number mapped Paired End	Paired end after filtering	Number Mapped Single End	Single End after filtering	Cove rage	CpG (%)	CHG (%)	CHH (%)
Buccal	6283213773	2174291340	516667527	1552616328	330217907	43	71.1	0.7	1.2
Blood	793730071	302431954	280560367	300859288	276523228	36	70.9	0.5	0.4

Table S3: GEO IDs for samples analysed

Sample	GEO ID
CD14+ 1	TBC
CD4+ 1	TBC
CD14+ 2	TBC
CD4+ 2	TBC
CD14+ 3	TBC
CD4+ 3	TBC
CD14+ 4	TBC
CD4+ 4	TBC
CD14+ 5	TBC
CD4+ 5	TBC
CD14+ 6	TBC
CD4+ 6	TBC
CD14+ 7	TBC
CD4+ 7	TBC
CD14+ 8	TBC
CD4+ 8	TBC
CD14+ 9	TBC
CD4+ 9	TBC
CD14+ 10	TBC
CD4+ 10	TBC
CD14+ 11	TBC
CD4+ 11	TBC
CD14+ 12	TBC
CD4+ 12	TBC
CD14+ 13	TBC
CD14+ 14	TBC
CD14+ 15	TBC
CD14+ 16	TBC
CD14+ 17	TBC
CD14+ 18	TBC
CD14+ 19	TBC
CD14+ 20	TBC
CD14+ 21	TBC
Buccal 1	TBC
Buccal 2	TBC
Buccal 3	TBC
CD34+ 1	TBC
CD34- 1	TBC
CD34+ 2	TBC
CD34- 2	TBC
Placenta 1	TBC
Placenta 2	TBC
Pancreas 1	TBC
Pancreas 2	TBC

Sperm 1	TBC
Sperm 2	TBC

Table S4: **Table with the TP and 1-FP rate for various different methylation cut-offs and region size cut-off. Highlighted in bold is the maximum harmonic mean found.** tDMRs were called using a Cochran-Mantel-Haenszel test (p-value<0.01) but it was necessary to filter these regions for further analysis. Therefore we investigated the effect of varying both region size and methylation difference cut-off on validation rates in sites also contained on 450K array. Here TP rate is defined as the number of regions overlapping probes called as a tDMP on the 450K array divided by the total number of probes called as tDMPs and the FP rate is defined as the number of regions called as tDMRs that overlapped with a 450K probe that were not called as a tDMP divided by the total number of tDMRs that overlapped with the 450K. tDMPs were obtained using dmpFinder in minfi. Supplementary Table 4 shows that a methylation difference cut-off of 30% and a minimum window size of 50bp (highlighted in bold) produced the highest harmonic mean of TP rate and 1-FP rate. Showing that over 80% of tDMPs were found and over 70% of tDMRs called were validated. A 50% methylation difference and 200bp window was chosen for main analysis as this allowed for a very low FP rate and hence our so-called tDMRs have a high validation rate.

Methylation Difference	Minimum Window Size	TP	1-FP	Harmonic Mean
0	0	0.942621101	0.494329042	0.648547186
0	50	0.939009526	0.500980353	0.653373098
0	100	0.932463546	0.505735189	0.655792021
0	150	0.924156923	0.509258818	0.656662333
0	200	0.91449596	0.512433022	0.656820254
0	300	0.891111011	0.517061645	0.654407431
0.1	0	0.940273577	0.498094928	0.651217679
0.1	50	0.936662002	0.504363052	0.655668969
0.1	100	0.930116022	0.509109294	0.65803555
0.1	150	0.921809399	0.512589065	0.658825884
0.1	200	0.912148436	0.515750564	0.658927655
0.1	300	0.888808632	0.520183074	0.65627527
0.2	0	0.91440567	0.563256545	0.697107868
0.2	50	0.910794095	0.569612225	0.700887917
0.2	100	0.904473839	0.57359916	0.702002451
0.2	150	0.896573518	0.575888944	0.701310614
0.2	200	0.887318857	0.578030621	0.70003433
0.2	300	0.865423683	0.58096206	0.695220245
0.3	0	0.81003115	0.687086669	0.743510761
0.3	**50**	**0.8070516**	**0.694379071**	**0.746487668**

0.3	100	0.801724527	0.697438968	0.745954566
0.3	150	0.794997968	0.698921837	0.743870505
0.3	200	0.78750395	0.700689098	0.74156432
0.3	300	0.768904338	0.703134134	0.73454994
0.4	0	0.605932012	0.790323875	0.685952396
0.4	50	0.603855356	0.797827448	0.687419973
0.4	100	0.600695228	0.800722739	0.686433797
0.4	150	0.595729312	0.801179232	0.683346029
0.4	200	0.590176516	0.802595416	0.680187409
0.4	300	0.578393752	0.803920997	0.672759778
0.5	0	0.302469414	0.860283026	0.447574727
0.5	50	0.301115074	0.868922086	0.447243125
0.5	100	0.299354431	0.872144764	0.445720152
0.5	150	0.296781184	0.872635845	0.442924795
0.5	200	0.293259898	0.873772364	0.439135065
0.5	300	0.286849352	0.874882703	0.43204375
0.6	0	0.058958963	0.875216638	0.11047573
0.6	50	0.058146359	0.898912058	0.109227321
0.6	100	0.05728861	0.908201305	0.10777863
0.6	150	0.055889125	0.908656145	0.105301427
0.6	200	0.054805652	0.91182266	0.10339659
0.6	300	0.052683852	0.915376677	0.099633376
0.7	0	0.00433389	0.711627907	0.008615312
0.7	50	0.004017877	0.832335329	0.00799715
0.7	100	0.00365672	0.857142857	0.007282372
0.7	150	0.003340707	0.854961832	0.006655408
0.7	200	0.003024694	0.881355932	0.006028699
0.7	300	0.002663537	0.911764706	0.005311557
0.8	0	0.000316013	0.319148936	0.0006314
0.8	50	0.000225723	0.555555556	0.000451264
0.8	100	0.000180579	0.642857143	0.000361056
0.8	150	0.000180579	0.666666667	0.00036106
0.8	200	0.000180579	0.727272727	0.000361068
0.8	300	0.000180579	1	0.000361092
0.9	0	0	0.208333333	0
0.9	50	0	0.5	0
0.9	100	0	1	0
0.9	150	0	1	0
0.9	200	0	1	0
0.9	300	0	1	0
1	0	0	0.090909091	0

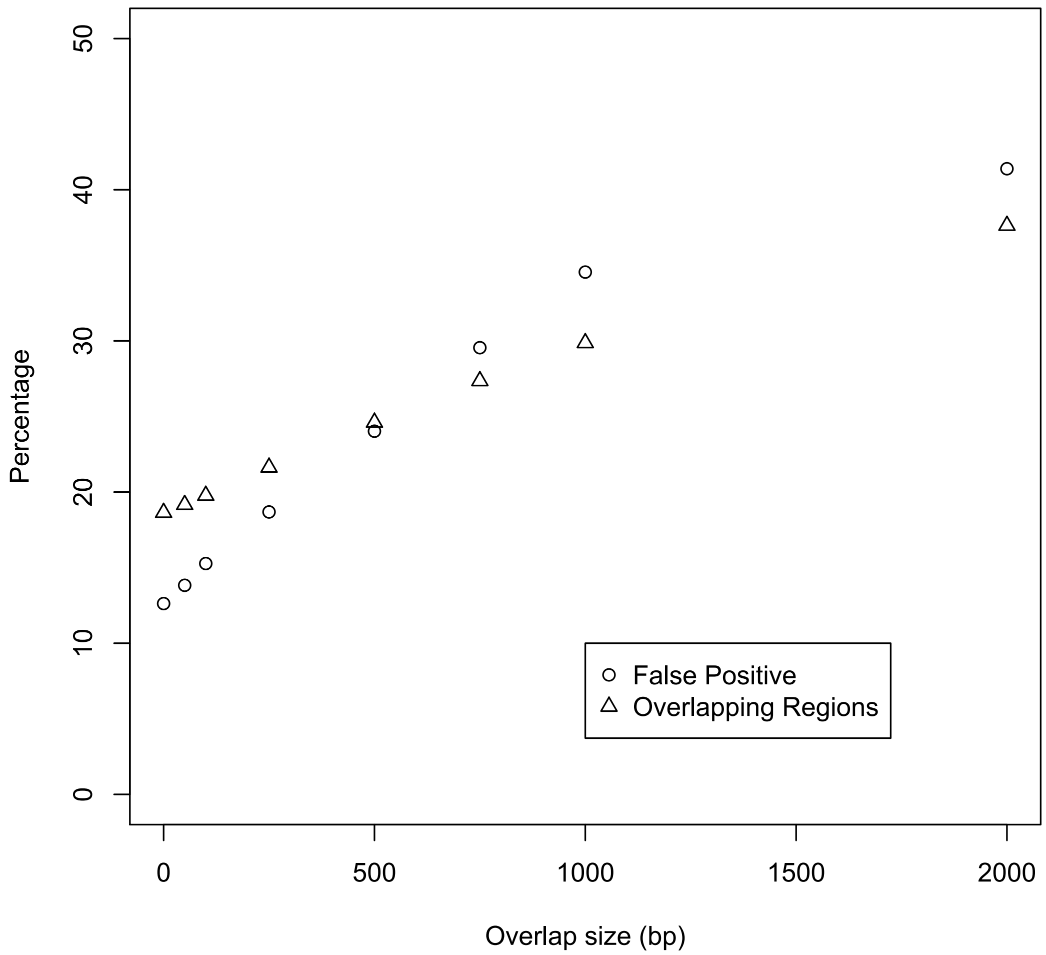

Figure S1: Analysis of varying the allowed overlap between BS-Seq tDMRs and 450K or RRBS-Seq data. As we increase the overlap the validation rate falls suggesting we are not capturing the true methylation state.

Location of External data used for paper excluding the >1000 450K samples that can be downloaded using Marmal-aid with the script supplied

Data Name	Location
UCSC Genes	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/knownGene.txt.gz
RegFeats	http://ftp.ensembl.org/RegulatoryFeatures_MultiCell.gff.gz
H3K27me3 PBMC	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM613877
H3K9me3 PBMC	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM613878
H3K9ac PBMC	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM613879

H3K36me3 PBMC	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM613880
H3K4me1 PBMC	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM613884
miRNA	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/wgRNA.txt.gz
Long non-coding RNA	ftp://ftp.sanger.ac.uk/pub/gencode/release_13/gencode.v13.long_noncoding_RNAs.gtf.gz
mRNA	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/all_mrna.txt.gz
Dnase Clustered	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClustered.bed.gz
Transcription Factor Chip	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClustered.bed.gz
CpG Islands	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/cpgIslandExt.txt.gz
H3K27ac ES	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM605307
H3K36me3 ES	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM605310
H3K27me3 ES	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM667622
H3K4me1 ES	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM667626
H3K4me2 ES	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM605314
H3K4me3 ES	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM605316
H2AZ ES	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM807391
Ctcf ES	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeOpenChromChip/wgEncodeOpenChromChipH1hescCtcfRawDataRep1.fastq.gz
Pol2 ES	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeOpenChromChip/wgEncodeOpenChromChipH1hescPol2RawDataRep1.fastq.gz
RRBS-Seq Skeletal Muscle	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/wgEncodeHaibMethylRrbsBcskeletalmuscle0111002BiochainSitesRep1.bed.gz
RRBS-Seq Islets	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/wgEncodeHaibMethylRrbsPanisletsSitesRep1.bed.gz
RRBS-Seq Brain	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/wgEncodeHaibMethylRrbsBcbrainh11058nBiochainSitesRep1.bed.gz
RRBS-Seq Kidney	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/wgEncodeHaibMethylRrbsBckidney0111

	002BiochainSitesRep1.bed.gz
RRBS-Seq Liver	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/wgEncodeHaibMethylRrbsBcliver0111002BiochainSitesRep1.bed.gz
DNaseI	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26328