PLOS | ONE

RESEARCH ARTICLE

Intra-Gene DNA Methylation Variability Is a Clinically Independent Prognostic Marker in Women's Cancers

Thomas E. Bartlett[1,2,3], Allison Jones[1], Ellen L. Goode[4], Brooke L. Fridley[5], Julie M. Cunningham[6], Els M. J. J. Berns[7], Elisabeth Wik[8], Helga B. Salvesen[8], Ben Davidson[9], Claes G. Trope[10], Sandrina Lambrechts[11], Ignace Vergote[11], Martin Widschwendter[1]*

1 Department of Women's Cancer, Elizabeth Garrett Anderson Institute for Women's Health, University College London, London, United Kingdom, **2** Deparment of Mathematics, University College London, London, United Kingdom, **3** CoMPLEX, University College London, London, United Kingdom, **4** Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, United States of America, **5** Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS, United States of America, **6** Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States of America, **7** Department of Medical Oncology, Erasmus MC-Cancer Center, Rotterdam, The Netherlands, **8** Department of Obstetrics and Gynaecology, Haukeland University Hospital, Bergen, Norway, **9** Department of Pathology, Oslo University Hospital, Norwegian Radium Hospital, University of Oslo, Faculty of Medicine, Institute of Clinical Medicine, Oslo, Norway, **10** Department of Gynaecological Oncology, Oslo University Hospital, Norwegian Radium Hospital, Oslo, Norway, **11** Division of Gynecologic Oncology, Department of Obstetrics and Gynecology and Leuven Cancer Institute, University Hospitals Leuven, Katholieke Universiteit Leuven, Leuven, Belgium

* M.Widschwendter@ucl.ac.uk

OPEN ACCESS

Citation: Bartlett TE, Jones A, Goode EL, Fridley BL, Cunningham JM, Berns EMJJ, et al. (2015) Intra-Gene DNA Methylation Variability Is a Clinically Independent Prognostic Marker in Women's Cancers. PLoS ONE 10(12): e0143178. doi:10.1371/journal.pone.0143178

Editor: Dajun Deng, Peking University Cancer Hospital and Institute, CHINA

Received: October 5, 2015

Accepted: October 30, 2015

Published: December 2, 2015

Data Availability Statement: DNA methylation data for the main OC data-set analysed here have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE72021.

Funding: This work was funded (MW, AJ) by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 305428 (Project EpiFemCare), by the National Institute for Health Research University College London Hospitals Biomedical Research Centre, and by the Eve Appeal and the European Network Translational Research in Gynaecological Oncology

Abstract

We introduce a novel per-gene measure of intra-gene DNA methylation variability (IGV) based on the Illumina Infinium HumanMethylation450 platform, which is prognostic independently of well-known predictors of clinical outcome. Using IGV, we derive a robust gene-panel prognostic signature for ovarian cancer (OC, $n = 221$), which validates in two independent data sets from Mayo Clinic ($n = 198$) and TCGA ($n = 358$), with significance of $p = 0.004$ in both sets. The OC prognostic signature gene-panel is comprised of four gene groups, which represent distinct biological processes. We show the IGV measurements of these gene groups are most likely a reflection of a mixture of intra-tumour heterogeneity and transcription factor (TF) binding/activity. IGV can be used to predict clinical outcome in patients individually, providing a surrogate read-out of hard-to-measure disease processes.

Introduction

Differences in DNA methylation (DNAm) levels are amongst the earliest changes in human carcinogenesis [1] and are a hallmark of cancer [2], offering the potential for novel strategies to predict cancer biology and outcome. The epigenetic differences which these changes give rise to are more stable than differences in gene expression level. Gene expression levels, as measured by RNA, are subject to periodic and transient variability (such as diurnal variation and

(ENTRIGO) of the European Society of Gynaecological Oncology (ESGO). TEB received funding from the UK Engineering and Physical Sciences Research Council (ESPRC) and the UK Medical Research Council (MRC) via UCL CoMPLEX. ELG received funding from the Fred C. and Katherine B. Andersen Foundation, NIH grants R01-CA122443, P50-CA136393 (the Mayo Clinic Ovarian Cancer SPORE) and P30-CA15083. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: BRCA, Breast cancer invasive carcinoma; DNAm, DNA methylation; EC, Endometrial cancer; ENCODE, Encyclopedia of DNA elements; FDR, False discovery rate; ITH, Intra-tumour heterogeneity; OC, Ovarian cancer; IGV, Intra gene variability of DNA methylation; TCGA, The Cancer Genome Atlas; TF, Transcription factor; UCEC, Uterine corpus endometrial carcinoma.

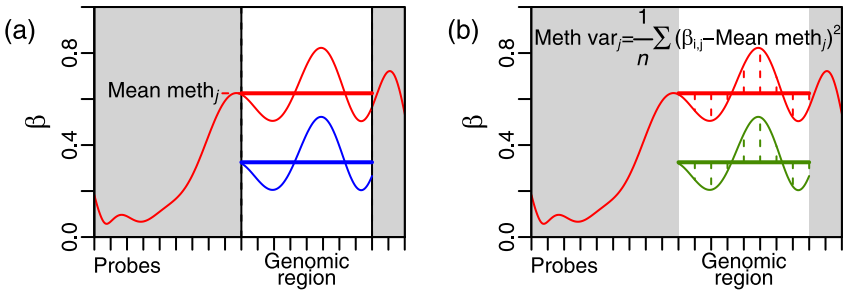


Fig 1. Per-gene methylation measures. (a) The mean methylation level over a specific genomic region is calculated separately for the TSS200 (promoter) and gene body genomic regions. The blue curve indicates the new position of the red curve after an additive global shift in methylation level, which might be due to technological or other experimental factors, and the difference between the horizontal red and blue lines (mean levels) illustrates the effect of this shift on the mean methylation level. (b) The intra-gene methylation variability (IGV) is calculated from the variation around the mean methylation level, i.e., from the dashed vertical lines, and is similarly calculated separately for the TSS200 and gene body genomic regions. The vertical green lines are changed very little compared to the vertical red lines, illustrating that such a global additive shift in mean methylation level has much less effect on IGV, which is therefore referred to as a 'self-calibrating measure'.

doi:10.1371/journal.pone.0143178.g001

mRNA instability), which do not apply to DNAm. Identifying reliable indicators of differences in DNAm patterns might provide a valuable lead for the development of DNA-based cancer biomarkers in tissue and bodily fluids.

Ovarian cancer (OC) and endometrial cancer (EC) are the most common gynaecological cancers [3]. Only one in three patients with advanced stage OC survive for five years after their initial diagnosis [4]. Very little is known about OC biology and how to manipulate this disease therapeutically. DNAm changes are important in cancer [5]; the epigenome is an interface between the genome and the environment [6, 7], and hence DNAm changes can measure exposure to environmental risk factors of cancer. DNAm biomarkers which represent a surrogate for patterns of gene interaction have previously been associated with clinical outcome in a wide variety of cancers [8], as well as specifically in women's cancers [9].

Sample to sample variability of DNAm at specific genomic locations is known to be important in the development of cancer [10, 11], and it has recently been shown that an increase in intra-gene variability of DNAm (IGV), a measure of within-sample methylation variability (Fig 1a), is highly associated with cancerous tissues in comparison to healthy [12]. Differential methylation is the commonly-used method by which methylation levels are compared between tissues, phenotypes and experimental conditions (equivalently to differential expression of genes). Here, we develop a prognostic signature based on IGV which is independent of well-known clinical prognostic features, and show that this IGV prognostic signature is likely a surrogate readout reflecting a mixture of intra-tumour heterogeneity and transcription factor (TF) binding/activity.

Results

Comparison of predictive robustness of per-gene methylation measures in data

To assess the effectiveness and robustness of IGV compared to mean methylation levels, we compared four per-gene methylation measures, based on mean methylation level and IGV (Fig 1). For each gene, we calculated mean methylation level and IGV, separately for the promoter (TSS200) and gene body regions, by using the Illumina Infinium HumanMethylation450 platform specifications of the CpGs in these regions for each gene. We considered different

genomic regions separately, because methylation patterns vary greatly from one genomic region to another, and the effect of methylation level on gene regulation varies according to genomic region. The four measures we compared, are as follows:

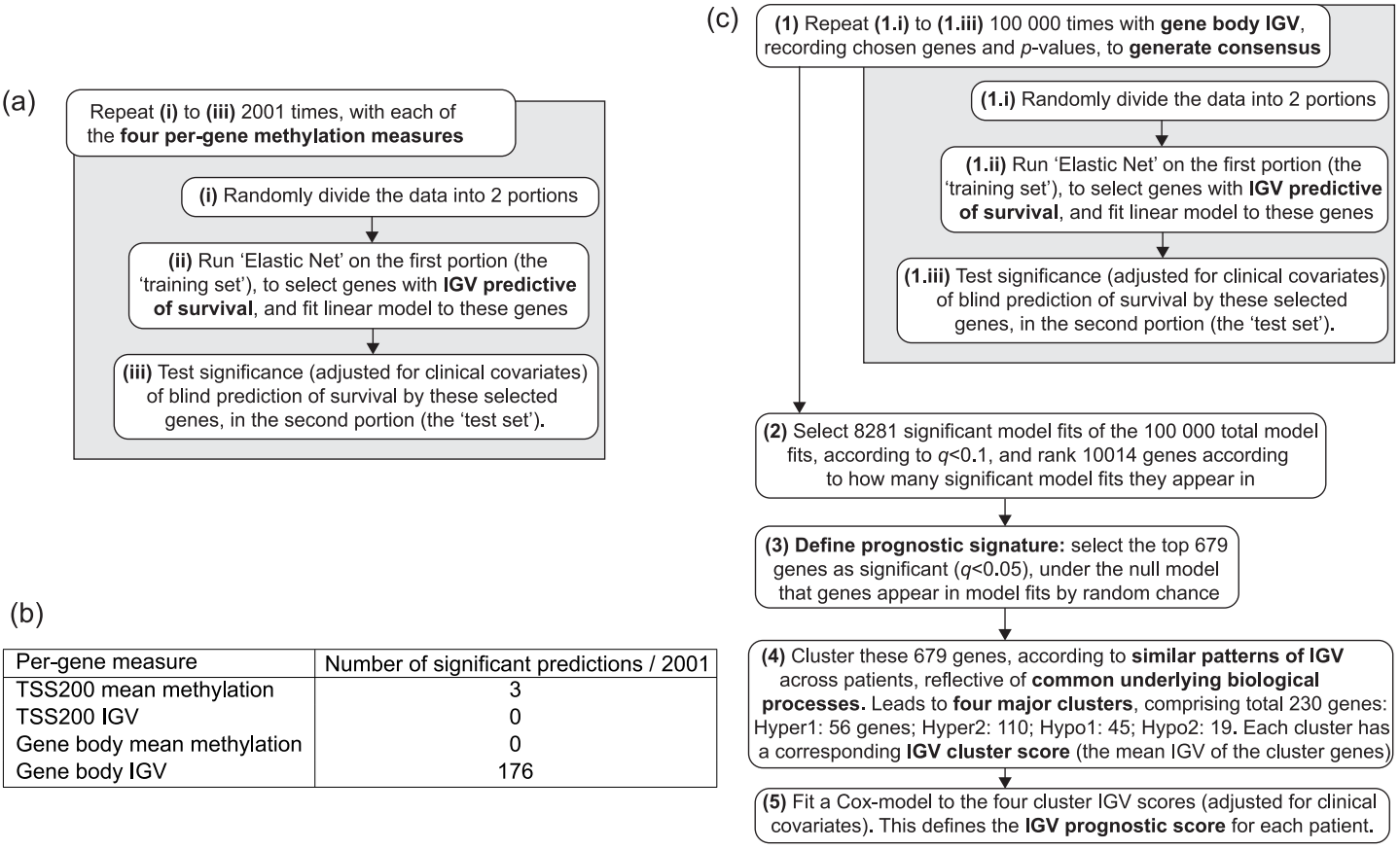
- TSS200 mean methylation
- TSS200 IGV
- Gene body mean methylation
- Gene body IGV

We obtained genome-wide DNAm profiles, via the Illumina Infinium HumanMethylation450 platform, from 218 primary OC samples. For each of the four measures described, we used 'Elastic Net' [13, 14] to find a prognostic selection of genes. Elastic net has been found to be an optimal linear modelling method to identify groups of genes which act together as part of a common biological process [15]. It is a regression method which 'chooses' the set of genes which model the data best, trying to include as few genes in the model as possible, whilst ensuring that the model predicts the outcome of interest as accurately as possible. In doing so, it discards genes which do not provide useful information, or which provide repeated information. As our aim is to find a minimal set of genes to use as a prognostic signature, it is important to note that amongst these genes, there will be groups of genes for which their IGV contains redundant or overlapping information, and there will be groups of genes for which IGV contains complementary information for each gene. Hence we chose to use the Elastic Net technique to accurately discern such a non-redundant grouping of genes as a minimal predictive set from very many possibilities, genome wide. We note that whilst this methodology may seem complex in this context, simpler methodology would not be able to discern these parsimonious groupings of genes in which overlapping and redundant information is kept to a minimum.

We assessed the effectiveness of the per-gene methylation measures as prognostic measures by randomly dividing the data into two portions: a 'training set', and a 'test set'. Elastic Net was used to select genes and fit a model to the training set, and the ability of this gene selection and model to blindly predict patient survival outcome (adjusted for clinical covariates) was assessed using the test-set. This was repeated 2001 times, and significantly predictive selected groups of genes were defined according to false discovery rate (FDR) adjusted [16] p-value (i.e., FDR q-value) < 0.1 (Fig 2a). As shown in Fig 2b, only gene body IGV predicts well.

Derivation of an ovarian cancer prognostic signature, and IGV prognostic score

We used IGV to derive an OC DNAm prognostic signature (Fig 2c), based on gene-body IGV (from here on simply referred to as 'IGV'). We did this by determining a consensus on a set of genes predictive of survival, by following the same procedure of splitting data into test and training sets, and then assessing the gene selection and fitted model for their ability to blindly predict patient survival outcome (adjusted for clinical covariates) in the test set. In order to ensure convergence to a stable result, we made 10^5 such partitions of the data, each resulting in a predictive selection of genes. Of these, 8281 were found as significant (FDR $q < 0.1$), and significance for each gene was then calculated based on the number of significant models in which that gene appeared. 679 genes were selected like this for inclusion in the OC prognostic signature at a significance level of FDR $q < 0.05$, with the least significant gene present in 1057 out of 8281 model fits. The top 100 most significant of these genes are shown in Supplementary Tables (S1 File).



Per-gene measure	Number of significant predictions / 2001
TSS200 mean methylation	3
TSS200 IGV	0
Gene body mean methylation	0
Gene body IGV	176

Fig 2. Overview of methods. (a) Methodology overview for comparison of the four per-gene methylation measures. (b) Results of this comparison. (c) Methodology overview for calculation of ovarian cancer IGV prognostic score.

doi:10.1371/journal.pone.0143178.g002

Genes often act together as part of biological pathways, and processes. Hence, we can expect that these 679 OC prognostic signature genes can be represented by a smaller number of underlying biological processes, which are important to disease progression. Grouping genes with similar experimental measurements by using clustering methodology is well established as an effective approach for determining clinically relevant prognostic markers [17, 18]. Hence, to uncover such groupings in the 679 genes of our OC prognostic signature, we carried out consensus clustering [19], to identify groups of genes with similar patterns of IGV across patients. Each cluster identified in this way reveals a different IGV trend, and therefore may correspond to a different underlying biological process, which gives rise to the pattern of IGV observed in that cluster. The clustering was carried out separately for genes which were individually associated with worse patient survival outcome for increased IGV ('hyper' genes) and for decreased IGV ('hypo' genes). The result was four clusters: two from the hyper genes, called clusters 'hyper 1' and 'hyper 2', and two from the hypo genes, called clusters 'hypo 1' and 'hypo 2'; they are shown in Supplementary Tables (S1 File). The mean IGV of the genes of each of the four clusters gives an IGV 'cluster score', for each cluster and for each patient, which are taken to be representative of the different IGV trends, and corresponding underlying biological processes, within the OC prognostic signature.

We then calculated an IGV prognostic score, by fitting a multivariate Cox proportional hazards model (accounting also for clinical covariates) to the four IGV cluster scores. It was not possible to fit such a model to the full set of 10014 genes, because there are many more

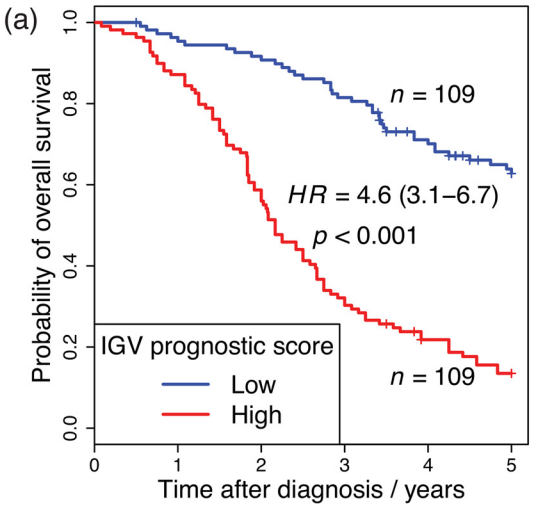
predictor variables (genes) than samples [20]. However, reducing the prognostic signature to 4 cluster scores, i.e., 4 predictors, allows the Cox proportional hazards model to be fitted. This results in a model coefficient for each cluster score/predictor; these are used to calculate the IGV prognostic score. The IGV prognostic score is a one-number prognostic indicator for a single sample/patient, and we note that it must be calculated based on all four cluster scores, to be significantly prognostic.

The median of this IGV prognostic score was used to divide the patients of the main OC data set into better and worse prognostic groups, shown in Fig 3a and 3b. The IGV prognostic score was validated in two independent sets of cancers derived from the Mullerian tract. A new OC set from the Mayo Clinic (n = 198) confirmed the prognostic capacity of the IGV prognostic score in both univariate (Fig 3c) and multivariate (Fig 3d) analyses. In order to test whether the IGV prognostic score is only limited to OC, or whether it is also predictive in other cancers which arise from the same embryological structure (i.e., the Mullerian duct), we applied our prognostic score to a publically available uterine corpus endometrioid carcinoma (UCEC) set from *The Cancer Genome Atlas* (TCGA) [21] (n = 358). Again, in both univariate (Fig 3e) and multivariate (Fig 3f) analyses, we were able validate the IGV prognostic score.

We note that using the median prognostic score from the main OC data-set (the training set) to dichotomise the patients of the Mayo OC and TCGA UCEC validation sets makes this a true assessment of the prognostic ability of this methodology. This is because by this method, the patients of the validation sets are classified one by one into a better or worse prognostic group, in terms of their DNAm measurements only. This classification is done according to a threshold or boundary dividing these prognostic groups (i.e., the median of the prognostic score in the training data-set), and this threshold is set entirely independently of these validation data-sets.

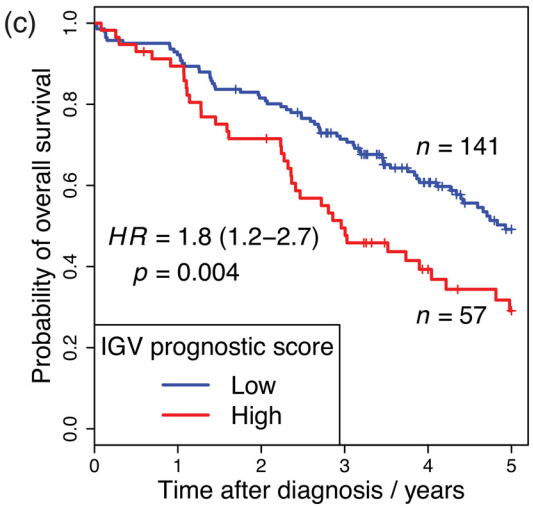
IGV and intra-tumour heterogeneity

We suggest that the IGV cluster scores are each representative of different biological processes, important for disease outcome. But what are these processes? To try to find some answers to this question, we first hypothesised that intra-tumour heterogeneity might be a reflection of IGV. The subject of intra-tumour heterogeneity is currently receiving a great deal of attention, uncovering much spatial and temporal diversity in genomic processes within individual tumours [22]. Ideally, the DNA methylome of individual cells from the same tumour sample should be analysed to address this question. As an alternative approach, we use here cross-sample methylation variance (i.e., mean methylation variance of individual CpGs of a specific gene-body region), as a measure of intra-tumour methylation heterogeneity, in order to assess how this varies as a function of IGV (Fig 4a). Cross-sample methylation variability is also a measure of how similar the methylation profiles are for the gene, across samples. If cross-sample methylation variability were a reflection of IGV, as IGV increases, we would expect to see a consistently increasing cross-sample methylation variance (Fig 4b, expected proportional fit). However, instead we see a pattern in which for low IGV, cross-sample methylation variance increases, whereas for high IGV, cross-sample methylation variance decreases again and is very low for the highest IGV values. In order to validate this further, we analysed two additional data sets, for which several samples from different regions of the same cancer have been taken. The first additional data-set is derived from endometrial cancers, where independent samples have been taken from 2 or 3 primary cancer and metastatic sites, in each of 10 patients (Fig 4c, one curve of best fit is shown per patient). The second is derived from prostate cancers, where 8 independent samples have been taken from the same tumour, from each of five cancer patients [23] (Fig 4d, one curve per patient). The pattern of these curves is almost identical to



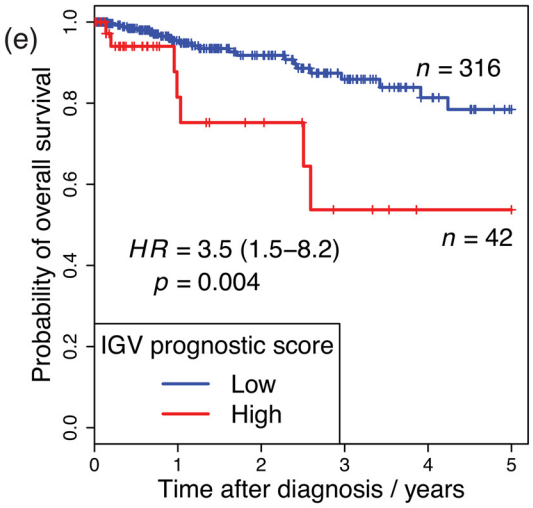
(b)

	HR (95%CI)	*p*	*n*
IGV prognostic score	4.1 (2.8-6.2)	< 0.001	209
Age	1 (0.7-1.4)	1	209
Stage	5.2 (1.9-14.4)	0.002	209
Grade	0.9 (0.6-1.4)	0.7	209
Residual disease	1.4 (1-2)	0.05	209



(d)

	HR (95%CI)	*p*	*n*
IGV prognostic score	1.9 (1.1-3.1)	0.01	149
Age	0.8 (0.5-1.3)	0.4	149
Stage	12.3 (1.5-99.2)	0.02	149
Grade	1.5 (0.5-4.2)	0.5	149
Residual disease	2.5 (1.4-4.4)	0.001	149



(f)

	HR (95%CI)	*p*	*n*
IGV prognostic score	2.8 (1.1-7.2)	0.04	308
Age	1.2 (0.5-2.9)	0.7	308
Stage	2.1 (0.8-5.5)	0.1	308
Grade	0.9 (0.3-2.9)	0.9	308
Residual disease	8.9 (3.1-26)	< 0.001	308

Fig 3. IGV OC prognostic signature validation. (a), (c) and (e): Comparison of survival curves of groups defined by the IGV prognostic score, in: (a) the main OC data set, (c) the Mayo Clinic OC validation set, (e) the uterine cancer TCGA validation set. The groups are divided by the median IGV prognostic score derived in the main OC DNAm data-set. The hazard ratio (*HR*) is displayed with 95% C.I. in brackets, with corresponding *p*-value calculated by univariate Cox regression. (d), (e) and (f): Multivariate Cox regression comparing the same groups defined by the IGV prognostic score.

doi:10.1371/journal.pone.0143178.g003

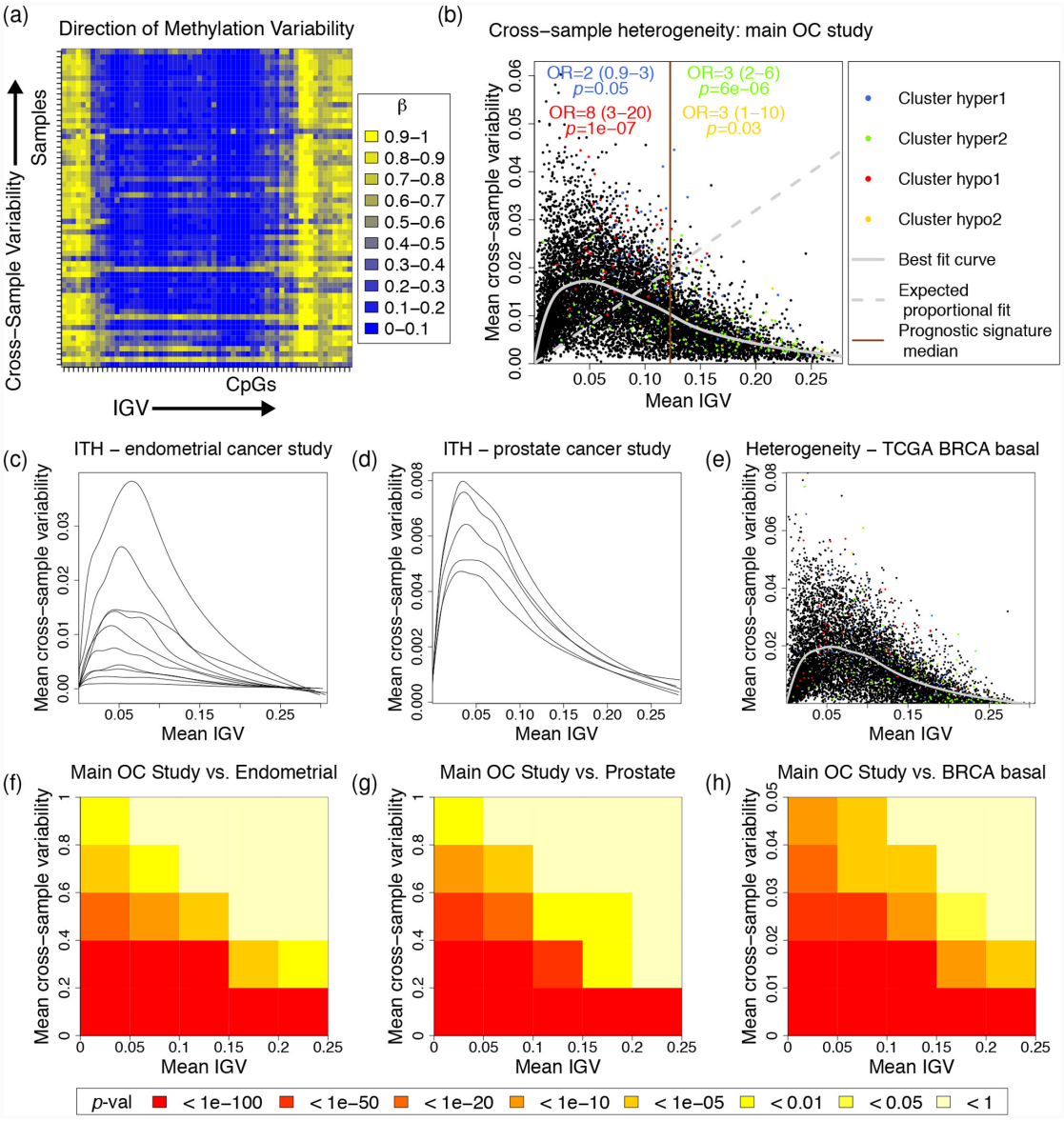


Fig 4. Comparison of IGV with Intra-Tumour Heterogeneity. (a) Cross-sample variability of methylation (Intra-tumour heterogeneity) and IGV are calculated in different and complementary directions. The heatmap displays the methylation profile of a single gene (horizontal axis), across multiple samples (vertical axis). (b)-(e) A characteristic pattern of high cross-sample variability (intra-tumour heterogeneity) when IGV is low, and vice-versa, is consistently observed across different studies: (b) Main OC data-set, (c) Endometrial cancer intra-tumour heterogeneity data-set, (d) prostate cancer intra-tumour heterogeneity data-set, (e) BRCA basal data-set. (f)-(h) The overlap of genes in each region of (b) with genes in equivalent regions of (c)-(e) is highly significant. In (c) and (d), each line relates to samples from a single patient, and is a best fit curve equivalent to that shown in (b) and (e). In (b), odds-ratios and *p*-values at the top of the plot show enrichment of the genes of each cluster, either side of the median IGV of the prognostic signature. Abbreviations: ITH (intra-tumour heterogeneity), OC (ovarian carcinoma), BRCA (breast cancer invasive carcinoma).

doi:10.1371/journal.pone.0143178.g004

the intra-tumour heterogeneity studies, in the main OC study which we used to identify the OC prognostic signature (Fig 4b), and in basal samples from the TCGA breast-cancer invasive carcinoma (BRCA) data-set (Fig 4e). The overlap of genes in all regions of these plots is also highly significant across data sets (Fig 4f–4h).

The genes of cluster hyper 1 are somewhat over-represented in the left half of Fig 4b, where IGV is lower, and cross-sample methylation heterogeneity is typically higher. This suggests

that the increased IGV of these genes is associated with intra-tumour heterogeneity. However, the genes of clusters hyper 2 and hypo 2 fall mostly in the region of high IGV and low cross-sample methylation variability (towards the right of Fig 4b). This means that, for the genes of these clusters, their methylation profiles tend to be similar in different samples from the same tumour, or from different tumours. In the case of cluster hyper 2, this corresponds to high methylation variability within a single gene in poor prognostic cases, and that this variability is consistently similar throughout the tumour and between tumours. Hence, the genes of cluster hyper 2 show high IGV in poor prognostic cases, yet appear to be independent of intra-tumour heterogeneity. Therefore, we speculate that the increased IGV of these genes is a tumour-cell inherent phenomenon, independent of intra-tumour heterogeneity. This means that the IGV prognostic signature combines measures of intra-tumour heterogeneity, with those of independent, tumour-cell inherent phenomena. We note that the terms 'hyper' and 'hypo', here relate to change, rather than absolute level. For example, S1 Fig shows that cluster hypo 2 has the highest IGV of any cluster; however, the IGV of this cluster is actually lower in poor compared to good prognostic cases.

The genes defining cluster hypo 1 have the highest mean cross-sample methylation variability (Fig 4), as well as the highest mean methylation level (S2 Fig), and the low IGV of the hypo 1 genes is associated with poor prognosis. At first, it seems difficult to explain that poor prognostic cancers have lower IGV in the hypo1 genes, yet these hypo1 genes also represent high sample-sample methylation heterogeneity. To explain this, we used a measure of CpG-CpG methylation variability, which we call the mean derivative [12], which is calculated as the average absolute difference in methylation levels between adjacent CpGs of the gene-body of a gene, in a single sample. The Illumina HumanMethylation 450K array measures the methylation levels of specific CpG loci, averaged across a mixed-up sample of many cells. Fig 5a and 5b shows two examples of how high methylation variability at the single-cell level might manifest in measurements acquired using this technology.

In the example of Fig 5a, we see that there is little cell-cell heterogeneity, although there is much variability within a gene. Hence, this results in measurements of high IGV, and low cross-sample methylation variability, as we see in cluster hyper 2. Then Fig 5b shows an example in which there is much cell-cell variability, as well as much variability within a gene. The result is that the cross-sample methylation variability of the array measurements is high, but because the highly variable methylation profiles 'average out' across the mixed-up cells in the sample, the net result is a measurement with low IGV. To examine whether this hypothesis is plausible, we use the mean derivative measure of CpG-CpG methylation variability (Fig 5c). By considering how heterogenous this CpG-CpG variability is across samples (Fig 5d), we are able to confirm that in the genes of cluster hypo 1, the CpG-CpG methylation variability tends to be more different across different cells than in any other cluster, as reflected by the high variance of the mean-derivative measurements. We are also able to confirm from Fig 5d that in the genes of cluster hyper 2, the CpG-CpG methylation variability tends to be less different across different cells than in any other cluster, as indicated by the low variance of the mean derivative. Hence, these data support the model shown in Fig 5a and 5b for genes in cluster hyper 2 and hypo 1, respectively.

Functional role of transcription-factor activity in IGV

As the genes comprising cluster hyper 2 seem to show the same IGV in most cells of the tumour, but the high IGV of the cluster hyper 2 genes is associated with poor prognosis, we deem the cluster hyper 2 IGV to be a 'consistent tumour-cell inherent phenomenon', which is likely to be regulated by differential binding of transcription factors (TF). Therefore, we

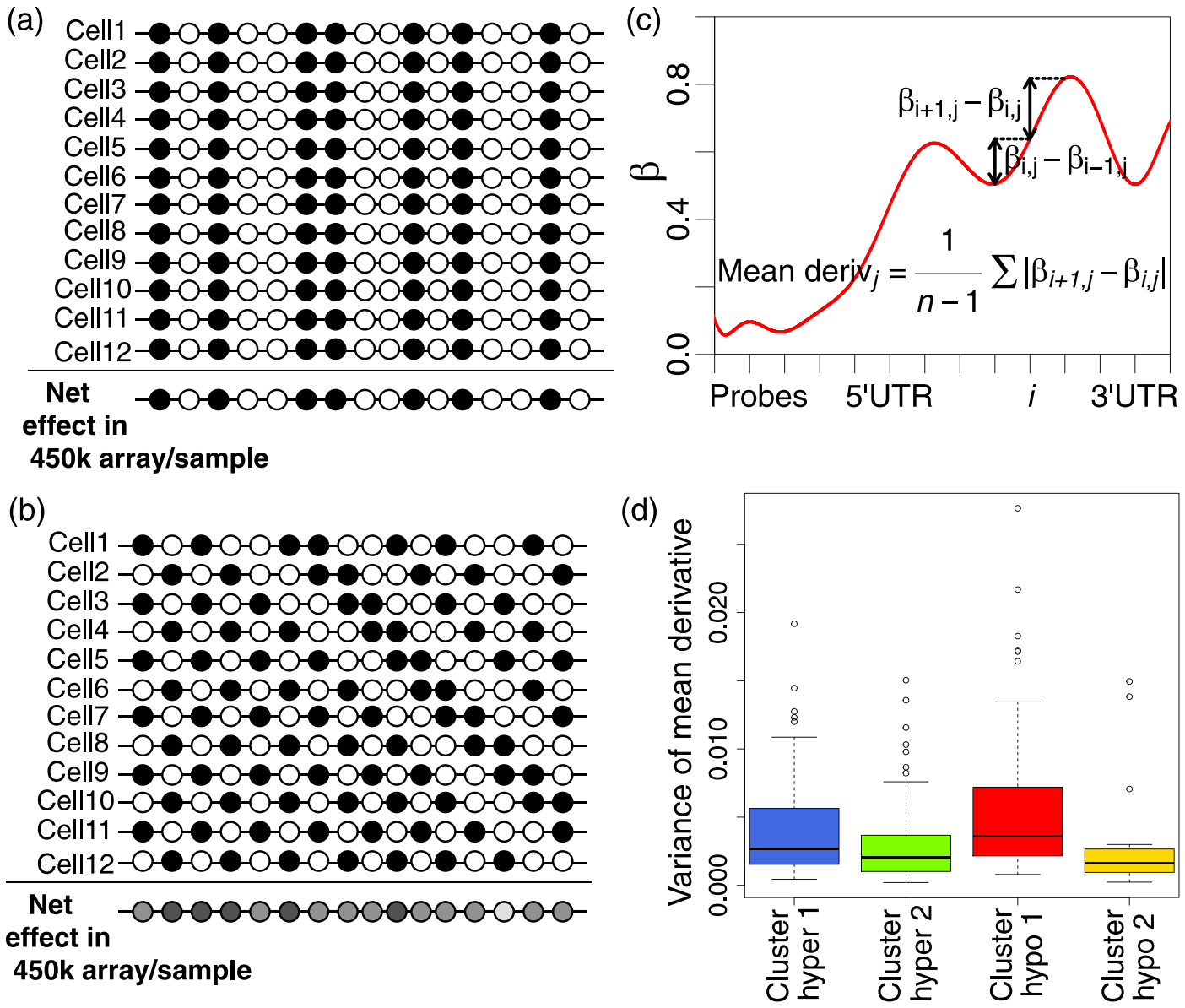


Fig 5. Heterogeneity and the effects of cell mixing on the 450K array. The 450K array provides methylation measurements from a mixed-up sample of multiple cells. (a) An example of a methylation pattern which is highly variable, in a similar way across cells. This leads to low cross-sample heterogeneity, and high IGV, as in cluster hyper 2. (b) An example of a methylation pattern which is highly variable, but in a heterogenous way across cells. This leads to high cross-sample heterogeneity, however the net effect of averaging the methylation profiles across the mixed up sample of many cells gives a measurement with low IGV, as in cluster hypo 1. (c) A measure of CpG-CpG methylation variability, calculated as the mean derivative, or the mean absolute difference in methylation level between adjacent CpGs. (d) The variability of the mean-derivative measure across samples quantifies the heterogeneity of the CpG-CpG methylation variability. Cluster hyper 2 is low according to (d), and hence corresponds to a pattern such as (a). Cluster hypo 1 is high according to (d), and hence corresponds to a pattern such as (b).

doi:10.1371/journal.pone.0143178.g005

examined TF binding to the gene body regions of the OC prognostic signature genes, and tested the correlation of TF expression with the IGV of the genes they bind to (in a TCGA set of basal breast cancers). We found that each prognostic signature cluster shows its own distinctive pattern of TF binding (Fig 6a), which we can hypothesise is associated with the biological processes responsible for the characteristic pattern of IGV observed in that cluster.

Transcription factor binding site information, obtained from the ENCODE (Encyclopedia of DNA Elements) project [24], was available for the gene body regions of all the genes represented on the Illumina HumanMethylation 450K array, for 55 transcription factors. We tested each of these 55 TFs, for significantly increased or decreased binding to the genes of each prognostic signature cluster. Cluster hypo 2 only consists of 19 genes, and hence we would not expect to see many significant correlations, due to small sample size. But interestingly, for cluster hyper 2 (comprised of genes whose methylation levels vary little across tumours but show higher IGV), we see that 20% (11/55) of the TFs tested show significantly more binding to these genes than expected, whereas 16% show significantly less binding than expected. For the gene clusters for which DNAm varies across/within tumours and have generally low IGV (clusters hyper 1 and hypo 1), not a single TF showed higher than expected binding, whereas 27% and 38% of TFs show lower than expected binding to the genes comprising cluster hyper 1 and hypo 1, respectively. This is consistent with the idea that TF binding is involved in distinct and different processes associated with IGV and methylation heterogeneity within a sample.

We also wanted to test the actual correlation of expression of the TFs with IGV of the genes they bind to, and genes they do not bind to, genome-wide. To do this, we used a TCGA set of basal breast cancers, for which 450k methylation data as well as expression data exist. We have already established a high degree of similarity in behaviour of our prognostic signature genes in OC and these TCGA BRCA basal samples (Fig 4). Further, it has been comprehensively demonstrated by the TCGA consortium that high-grade serous ovarian and uterine and BRCA basal cancers are extremely molecularly similar [25]. Fig 6b and 6c show TFs with significantly more positive, and more negative, correlation with IGV of the genes they bind to, compared to the genes they do not. It is interesting that the two most highly ranked transcription factors according to increased positive correlation of their expression with IGV in bound genes, *Rad21* and *Brg1* (*SMARCA4*), are both parts of chromatin modifying complexes with relevance to stem cell identity [26, 27]. In particular, *Brg1* (*SMARCA4*) has been shown recently to have particular relevance to small-cell ovarian cancer [28–30]. The overlap between the TFs which show significantly different binding patterns in relation to the OC prognostic signature genes, and TFs which display significantly altered correlation of their expression with IGV of genes they bind to, is shown in Fig 6d. Much relevant detail has already been reported about most of these TFs (references noted in the figure): either their binding is influenced by methylation (or *vice-versa*), or they are involved with chromatin remodelling in stem cells. The TFs shown in Fig 6d are important to the processes underlying disease progression, which are associated with our OC prognostic signature (TFs with known relevance are indicated with a reference to the relevant study [26, 31–40]). Therefore we hypothesise that IGV, in our OC prognostic signature gene panel, represents a surrogate measure for their activity and role in disease transformation.

Association of prognostic signature CpGs with CpG islands and enhancer regions

The location of CpGs relative to CpG islands (CGIs) is known to be an important determinant of the functional role of these CpGs [41]. We tested for enrichment of probes annotated to the CGI regions 'island', 'shore' and 'shelf' amongst all gene body annotated probes, as well as probes annotated to gene bodies of the genes of our prognostic signature, and of the four clusters. While we found that gene body probes were overall significantly depleted for probes in these CGI regions, the opposite was true for gene bodies of our prognostic signature (see Supplementary Tables in S1 File). This effect appears to be largely driven by the second cluster. This indicates a prominent role for CpG islands in the relevant areas of the genes of our prognostic signature.

(a)

	Hyper1	Hyper2	Hypo1	Hypo2
BAF155	q=0.00053, OR=0.37 (0.2-0.67)	q=5.3e-05, OR=2.5 (1.6-4)	q=1.5e-06, OR=0.18 (0.07-0.38)	q=0.65, OR=1.4 (0.49-4.1)
BAF170	q=0.053, OR=0.53 (0.27-0.98)	q=0.0089, OR=1.8 (1.2-2.6)	q=0.0075, OR=0.32 (0.13-0.7)	q=0.64, OR=1.3 (0.47-3.6)
BCL3	q=0.56, OR=0.8 (0.41-1.5)	q=0.01, OR=0.48 (0.28-0.79)	q=0.29, OR=0.56 (0.24-1.2)	q=0.56, OR=0.58 (0.14-1.8)
c-Fos	q=0.89, OR=1 (0.57-1.9)	q=0.0013, OR=0.42 (0.24-0.7)	q=0.0037, OR=0.26 (0.08-0.66)	q=0.29, OR=1.8 (0.66-5)
c-Myc	q=0.00018, OR=0.34 (0.18-0.62)	q=2.6e-07, OR=3.3 (2-5.4)	q=5.6e-05, OR=0.25 (0.11-0.52)	q=0.49, OR=1.5 (0.53-4.4)
CEBPB	q=0.16, OR=0.59 (0.32-1.1)	q=0.7, OR=0.91 (0.61-1.4)	q=0.0043, OR=0.3 (0.12-0.66)	q=0.7, OR=1.2 (0.44-3.4)
CTCF	q=0.34, OR=0.7 (0.38-1.2)	q=1.6e-05, OR=0.37 (0.23-0.58)	q=0.54, OR=0.79 (0.41-1.5)	q=0.34, OR=1.7 (0.64-5)
EBF	q=0.088, OR=0.57 (0.31-1)	q=0.44, OR=0.85 (0.57-1.3)	q=0.041, OR=0.44 (0.22-0.86)	q=0.14, OR=0.44 (0.14-1.2)
FOSL2	q=0.04, OR=0.5 (0.25-0.93)	q=0.12, OR=0.72 (0.47-1.1)	q=3e-05, OR=0.15 (0.039-0.41)	q=0.12, OR=0.4 (0.096-1.2)
FOXP2	q=0.12, OR=0.19 (0.0047-1.1)	q=0.023, OR=0.19 (0.023-0.72)	q=0.06, OR=0 (0-0.92)	q=1, OR=0.58 (0.014-3.7)
GABP	q=0.012, OR=0.26 (0.051-0.79)	q=0.0041, OR=2.1 (1.3-3.1)	q=0.0054, OR=0.11 (0.0026-0.62)	q=0.56, OR=0.53 (0.06-2.3)
GR	q=0.052, OR=0.49 (0.21-1)	q=0.029, OR=0.57 (0.33-0.93)	q=0.029, OR=0.33 (0.1-0.83)	q=0.029, OR=0.14 (0.0034-0.9)
HEY1	q=0.029, OR=0.52 (0.29-0.93)	q=2.9e-11, OR=4.6 (2.8-8)	q=2.5e-06, OR=0.18 (0.067-0.4)	q=0.65, OR=1.3 (0.48-3.7)
HNF4A	q=0.023, OR=0.44 (0.2-0.86)	q=0.15, OR=0.7 (0.44-1.1)	q=0.00039, OR=0.18 (0.046-0.49)	q=0.48, OR=0.64 (0.18-1.9)
Ini1	q=7.6e-05, OR=0.33 (0.19-0.58)	q=3.1e-08, OR=5.1 (2.6-11)	q=3.1e-08, OR=0.17 (0.078-0.33)	q=0.21, OR=2.4 (0.68-13)
JunD	q=0.3, OR=0.71 (0.4-1.2)	q=0.0058, OR=0.56 (0.37-0.84)	q=0.0054, OR=0.35 (0.17-0.71)	q=1, OR=1.1 (0.38-2.9)
Max	q=0.0022, OR=0.42 (0.23-0.74)	q=1e-04, OR=2.5 (1.6-4)	q=0.00018, OR=0.28 (0.13-0.56)	q=0.36, OR=1.6 (0.58-5.3)
NFKB	q=0.008, OR=0.46 (0.26-0.81)	q=1.4e-07, OR=5.6 (2.6-14)	q=0.007, OR=0.41 (0.21-0.77)	q=0.31, OR=2 (0.57-11)
NRSF	q=0.0091, OR=0.44 (0.22-0.81)	q=0.25, OR=0.8 (0.53-1.2)	q=8e-04, OR=0.26 (0.11-0.58)	q=0.15, OR=0.43 (0.12-1.3)
Pbx3	q=1, OR=0.88 (0.46-1.6)	q=0.038, OR=0.55 (0.33-0.88)	q=0.29, OR=0.57 (0.24-1.2)	q=1, OR=1 (0.32-2.9)
POU2F2	q=0.0035, OR=0.39 (0.2-0.73)	q=4e-04, OR=2.2 (1.4-3.3)	q=0.3, OR=0.67 (0.34-1.3)	q=0.65, OR=1.3 (0.48-3.7)
PU.1	q=0.46, OR=1.3 (0.76-2.5)	q=0.083, OR=0.67 (0.45-0.99)	q=0.023, OR=0.42 (0.21-0.81)	q=0.82, OR=0.83 (0.3-2.3)
Rad21	q=0.84, OR=0.79 (0.44-1.4)	q=1.7e-05, OR=0.37 (0.23-0.59)	q=0.88, OR=0.93 (0.48-1.8)	q=0.86, OR=1.4 (0.5-3.8)
Sin3Ak-20	q=0.034, OR=0.39 (0.14-0.92)	q=0.0023, OR=2 (1.4-3.1)	q=0.013, OR=0.24 (0.047-0.75)	q=1, OR=0.88 (0.21-2.8)
STAT1	q=0.0064, OR=0.37 (0.18-0.72)	q=0.38, OR=1.2 (0.8-1.8)	q=0.044, OR=0.46 (0.21-0.93)	q=0.23, OR=1.9 (0.7-5.4)
TAF1	q=0.34, OR=0.76 (0.43-1.3)	q=3.8e-12, OR=5.8 (3.2-11)	q=0.042, OR=0.48 (0.24-0.91)	q=0.084, OR=2.9 (0.91-12)
TCF12	q=0.0021, OR=0.32 (0.14-0.65)	q=0.92, OR=1 (0.69-1.5)	q=0.33, OR=0.62 (0.3-1.2)	q=0.92, OR=0.87 (0.29-2.4)
USF-1	q=0.024, OR=0.47 (0.23-0.91)	q=0.0078, OR=0.51 (0.31-0.8)	q=0.0084, OR=0.33 (0.12-0.74)	q=1, OR=1 (0.34-2.8)

(b)

	Median bound	Median unbound cor	q-val
Rad21	0.09	0.035	1.1e-21
Brg1	0.12	0.076	1.6e-12
GABP	0.072	0.041	3.9e-10
c-Myc	0.043	0.025	4.6e-06
Nrf1	0.11	0.086	4.7e-06
BCL11A	0.069	0.049	1.7e-05
FOXP2	0.083	0.04	1e-04
Pbx3	0.016	0.0027	0.00038
CTCF	0.02	0.01	0.001
SRF	0.037	0.0075	0.005
SIX5	0.023	-0.0039	0.0076
HNF4A	0.015	0.0059	0.014
Sin3Ak-20	0.11	0.096	0.026

(c)

	Median bound	Median unbound cor	q-val
GR	-0.22	-0.11	2.5e-51
BCL3	-0.19	-0.11	5.5e-28
PU.1	-0.18	-0.097	5.9e-27
NRSF	-0.11	-0.062	1.8e-21
c-Jun	-0.063	-0.037	3.7e-18
c-Fos	-0.095	-0.048	2e-17
BATF	-0.23	-0.14	4.1e-17
JunD	-0.11	-0.067	5.6e-12
TCF12	-0.029	-0.0056	4e-11
RXRA	-0.083	-0.049	1.2e-10
EBF	-0.12	-0.093	1.9e-08
p300	-0.06	-0.042	3e-08
FOSL2	-0.11	-0.073	7.5e-08
STAT1	-0.12	-0.092	4.1e-06
IRF4	-0.13	-0.12	0.00012
NFKB	-0.024	-0.013	0.0039

(d)

	Hyper1	Hyper2	Hypo1	Hypo2
Increased Binding, Positive Correlation with IGV		c-Myc [31] GABP [32] Sin3Ak-20 [33]		
Increased Binding, Negative Correlation with IGV		NFKB [34]		
Decreased Binding, Positive Correlation with IGV	c-Myc [31] GABP [32] HNF4A Sin3Ak-20 [33]	CTCF [26] FOXP2 [35] Pbx3 Rad21 [26]	c-Myc [31] GABP [32] HNF4A Sin3Ak-20 [33]	
Decreased Binding, Negative Correlation with IGV	FOSL2 NFKB [34] NRSF [36] STAT1 TCF12	BCL3 c-Fos [37] GR JunD [38]	c-Fos [37] EBF [39] FOSL2 GR JunD [38] NFKB [34] NRSF [36] PU.1 [40] STAT1	GR

Fig 6. Transcription Factor Binding and Expression Correlation with IGV. (a) False discovery rate adjusted p-values and odds-ratios (OR) show enrichment of binding of specific transcription factors (TFs), to the gene body regions of the genes of each cluster. TFs for which binding is significantly over or under enriched (Fisher's exact test, FDR $q < 0.05$) are coloured green and red, respectively. (b) TFs which show significantly more positive correlation with IGV of the genes they bind to, compared to the genes they do not bind to. (c) TFs which show significantly more negative correlation with IGV of the genes they bind to, compared to the genes they do not bind to. (d) TFs which are significant according to (a) and either (b) or (c); TFs with known relevance are indicated with a reference to the relevant study. The lack of enrichment of TF binding to the genes of cluster hypo2, is a reflection of the small number (19) of genes in this cluster.

doi:10.1371/journal.pone.0143178.g006

Location of CpGs relative to enhancer regions is also known to be relevant to the functional role of CpGs. We tested whether there was enrichment of methylation sites annotated to enhancers in gene bodies in general, finding that there is, as would be expected. Then, we tested enhancer enrichment similarly in the prognostic signature gene bodies, and the gene bodies of the individual clusters. We found that there is an even greater enrichment in the prognostic signature gene bodies than in gene bodies in general, which is consistent with IGV being mediated by transcription factor binding. This effect seems to be driven particularly by the 'hypo' clusters, for which methylation variability decreases with worse prognosis. These results are shown in Supplementary Tables (S1 File).

Discussion

We have found that IGV (a per-gene measure of intra-gene variability of DNAm) is a far more robust prognostic marker tool than mean methylation levels: Fig 2b indicates that gene body IGV has the potential to become an effective prognostic tool. While it is true that the Illumina HumanMethylation 450K array provides more DNAm measurements for the gene-body than for any other genomic region, and hence gene-body derived measures can potentially provide more information than those derived from the promoter region when using this technology, this is unlikely to be the whole explanation for its effectiveness in this study.

We note that it has previously been found that the most variably methylated CpGs occur more frequently in gene bodies than in promoters [42]. However, while it is well established that promoter methylation in CpG-dense regions is associated with gene repression [41], the effects of gene-body methylation are less clear. Gene body methylation has recently been shown to have a direct effect on gene expression level [43], however it may also be associated with other influences on transcription and translation, such as prevalence of alternatively spliced gene products [41]. Findings are also starting to emerge that gene-body methylation may be an effective therapeutic target in cancer [43].

The OC prognostic signature we have developed based on IGV, is able to blindly predict patient prognostic outcome in two independent data sets from studies by the Mayo Clinic and TCGA (n = 198 and n = 358, respectively), with highly statistically significantly different clinical outcomes between these groups (p = 0.004 in both data sets). The methodology we present here is, after calibration on a training data-set, able to classify patients one by one without reference to any more new samples into better and worse prognostic groups. Thus, our method gives a prediction of better or worse prognosis individually to patients. For this reason, it can be considered to be a true prognostic measure.

It is becoming increasingly clear that understanding intra-tumour heterogeneity, is crucial to understanding cancer biology [22, 23], including ovarian cancer [44], and recent work has shown the effectiveness of intra-tumour heterogeneity as a prognostic marker [45]. Asking the question, what is IGV, we examined whether intra-tumour methylation heterogeneity might be a reflection of IGV, finding that while for genes with relatively low IGV this may be true, for genes with high IGV, intra-tumour methylation heterogeneity does not appear to reflect IGV. Therefore, we may hypothesise that in these genes, IGV represents a tumour-cell inherent phenomenon. Investigating further the reasons for this phenomenon, by looking at binding of TFs and the correlation of their expression with IGV of genes they bind to, revealed a distinctive pattern of TF binding to different groups of genes, and identified a panel of TFs which are highly associated with prognostic IGV. However, the TF binding maps we analysed here is not exhaustive, and so this picture can be expected to become fuller, as more such TF binding data become available. We have also found evidence of the importance of CpG islands to the functional role of IGV in the genes of our prognostic signature and clusters.

Cancer is a heterogenous disease, which can, even within the same tissue type, show very different molecular characteristics between patients. Hence, it is becoming clear that for our mechanistic understanding of cancer to progress, we must focus on large-scale data-sets (i.e, 'big-data'), which are able to capture such heterogeneity with sufficient statistical power [45, 46]. Such analyses require computational statistical tools which are relatively new to medical science, which in turn requires interdisciplinary collaboration. In this study, we have made use of several such tools, to derive our prognostic signature gene-panel, and then to identify common molecular patterns within this gene-panel, which reflect heterogenous oncogenic processes. The methodology we present here is computationally efficient, and would naturally scale well to larger data-sets, and would be applicable to analysis of cancer data from a wide range of tissues of origin.

We have conclusively demonstrated that our OC prognostic signature is an effective and robust prognostic tool, and we also hypothesise that it is an easy to measure surrogate for disease processes mediated by specific transcription factors. IGV is a robust prognostic marker, which is independent of known clinical prognostic factors.

Methods

Data and preprocessing

The main ovarian cancer (OC) data set, which was used to derive the OC prognostic signature, consists of 221 samples each of which was taken from a different patient, of whom 158 died from the disease before the end of the study. For each sample, a DNA methylation profile collected via the Illumina Infinium HumanMethylation450 platform was available, together with information on the clinical variables survival status (alive or not), survival time (i.e., time to last follow up or time to death), disease stage (I-IV), disease grade (1–3), and residual disease status (present or not). 3 samples were removed due to missing clinical data, leaving the the $n = 218$ samples used to derive the OC prognostic signature. A further 9 samples were excluded from the multivariate analysis of the IGV prognostic score, due to additional missing clinical data.

An independent data set from a study of OC carried out by the Mayo Clinic was used for validation of the OC prognostic signature. Data from this study similarly included a DNA methylation profile for each sample collected via the Illumina Infinium HumanMethylation450 platform; clinical data was also available for this data set for the same variables as the main OC data set. There were $n = 198$ samples in this data set, of whom 115 died from the disease before the end of the study. 49 samples were excluded from the multivariate analysis of the IGV prognostic score, due to missing clinical data.

An additional independent data set from a study of uterine corpus endometrioid carcinoma (UCEC) for further validation of the OC prognostic signature was downloaded with the *The Cancer Genome Atlas* (TCGA) project [21]. Data from this study similarly included a DNA methylation profile for each sample collected via the Illumina Infinium HumanMethylation450 platform, which was downloaded at level 3; clinical data was also downloaded if possible for each sample for the same variables as the OC data set. There were 358 samples in this data set, of whom 32 died from the disease before the end of the study. 50 samples were excluded from the multivariate analysis of the IGV prognostic score, due to missing clinical data.

For the intra-tumour heterogeneity analysis, we considered two data sets, one from endometrial cancer (EC) (samples from multiple metastatic sites for each of 10 patients), and one from prostate cancer [23] (multiple samples from the same tumour for each of 5 patients). For comparison with cross-patient methylation heterogeneity, we downloaded DNAm data for breast cancer invasive carcinoma (BRCA) basal samples from TCGA (42 samples). Each of these data-sets included a DNA methylation profile for each sample collected via the Illumina

Table 1. Data-sets analysed.

Data-set	Patients	Samples per patient	Deaths	Removed
Main OC DNAm	221	NA	158	12
Mayo OC DNAm	198	NA	115	49
TCGA UCEC DNAm	358	NA	32	50
Endometrial ITH DNAm	10	2–3	NA	NA
Prostate ITH DNAm	5	16	NA	NA
TCGA BRCA basal DNAm	42	NA	NA	NA
TCGA BRCA basal Expr	42	NA	NA	NA

Abbreviations: ITH, intra-tumour heterogeneity; DNAm, DNA methylation; OC, ovarian cancer; UCEC, uterine corpus endometrial carcinoma; BRCA, breast cancer invasive carcinoma.

doi:10.1371/journal.pone.0143178.t001

Infinium HumanMethylation450 platform, again downloaded at level 3 for the TCGA BRCA data-set. For the gene expression analysis in BRCA basal samples, we downloaded gene expression data for the same 42 samples from TCGA, at level 3.

Probes with non-unique mappings and which map to SNPs had already been removed from the UCEC and BRCA TCGA DNAm data before they were downloaded, and these same probes were also removed from the other DNAm data sets. Probes mapping to sex chromosomes were also removed (by prior agreement); in total 98384 probes were removed from the DNAm data sets, of the 482421 probes originally present on the array. After removal of these probes, 270985 probes with known gene annotations remained. Individually for each data set, probes were then removed if they had less than 95% coverage across samples; probe values were also replaced if they had corresponding detection p-value greater than 5%, by KNN (k nearest neighbour) imputation ($k = 5$).

A summary of the data-sets analysed here appears in Table 1. A detailed summaries of the patient cohorts of the main ovarian and uterine cancer DNA methylation data-sets analysed here appear in Table 2.

Per-gene methylation measures

Four per-gene measures were tested, as follows:

- **TSS200 mean** The mean methylation level of the probes annotated to the TSS200 region, which is the region within 200bp upstream of the TSS (transcriptional start site); approximately the promoter region.
- **TSS200 IGV** The variance of the methylation level of the probes annotated to the TSS200 region.
- **Gene body mean** The mean methylation level of the probes annotated to the gene body.
- **Gene body IGV** The variance of the methylation level of the probes annotated to the gene body.

To calculate these measures, annotation information specifying which probes map to each gene and genomic region was used, as downloaded from Gene Expression Omnibus (GEO) [47], and as part of the *R* / *Bioconductor* software package *IlluminaHumanMethylation450k*. The mean methylation was calculated for genes with any number of probes annotated to the relevant genomic region (12970 and 15839 genes for TSS200 and gene body respectively). The

Table 2. Patient cohort details of the main DNA methylation data-sets analysed.

Data-set	Total patients	Stage 3–4	Grade 3	Age 60 or over	Residual disease
Main OC	221	183 (83%)	144 (65%)	94 (43%)	92 (42%)
Mayo OC	198	158 (80%)	164 (83%)	114 (58%)	70 (35%)
TCGA UCEC	358	103 (28%)	226 (63%)	241 (67%)	47 (18%)

doi:10.1371/journal.pone.0143178.t002

methylation variance was calculated for genes with at least 3 probes annotated to the relevant genomic region (7557 and 10014 genes for TSS200 and gene body respectively).

Cross-validation to compare per-gene methylation measures and derive OC prognostic signature

The samples (patients) of the main OC data-set were randomly split in to a 'training set' (2/3 of the data, 145 samples) and a 'test set' (the remaining 1/3 of the data, 73 samples). The Elastic Net [13, 14] was used to select a prognostic group of genes and fit a predictive model to these genes based on the training set; this model was then assessed using the test set. This was repeated 2001 times for each of the four per-gene methylation measures.

As the aim here is to predict clinical outcome, the Elastic Net was used in its penalised Cox regression form, as implemented in the *R* package *GLMNET* [14]. Cox regression fits the model by setting the model coefficients so as to maximise the partial likelihood, as defined by Eq (1),

$$L(\boldsymbol{\theta}) = \prod_{j \in S} \frac{e^{\boldsymbol{\theta}^\top \mathbf{x}_j}}{\sum_{j' \in R_j} e^{\boldsymbol{\theta}^\top \mathbf{x}_{j'}}}, \tag{1}$$

where $\boldsymbol{\theta}$ denotes the vector of model coefficients, $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ are the vectors of predictor variable values for samples j and j' respectively (here, per-gene methylation measures), S is the set of patients who died during the study, and R_j is the set of samples 'at risk' during the time interval when patient j died, defined as $R_j = \{j' | Y_{j'} \geq Y_j\}$, where Y_j and $Y_{j'}$ are the times of death of patients j and j' respectively. The Elastic Net penalises the log-likelihood corresponding to Eq (1), constraining it according to the magnitude of the model fit coefficients, by subtracting this constraint from the likelihood; in doing so, it 'chooses' the best combination of predictor variables (per-gene methylation measures), by adjusting the corresponding model coefficients, and setting these coefficients to zero where the variables provide no useful information or redundant information. The constraint is a combination of some multiples of the L_1 and L_2 norms of the model fit coefficients; the severity and balance of the constraint is controlled by the parameters λ (a 'magnitude' parameter) and α (a 'blending' parameter). Hence, the Elastic Net Cox model is fitted by finding model coefficients $\hat{\boldsymbol{\theta}}$ which maximise the penalised log likelihood $\phi(\boldsymbol{\theta}, \lambda, \alpha)$ in Eq (2),

$$\phi(\boldsymbol{\theta}, \lambda, \alpha) = \frac{2}{N} l(\boldsymbol{\theta}) - \lambda \left(\alpha \parallel \boldsymbol{\theta} \parallel_{L_1} + \frac{(1-\alpha)}{2} \parallel \boldsymbol{\theta} \parallel^2_{L_2} \right), \tag{2}$$

where N is the number of samples, $\| \cdot \|_{L_1}$ and $\| \cdot \|_{L_2}$ are the L_1 and L_2 norms, and $l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$. The *R* package *GLMNET* used for these model fits sets the λ parameter internally using ten-fold cross validation, and requires the user to set the α parameter ($0 \leq \alpha \leq 1$), which was in this case set by choosing the value which minimises the model error after trialling values from 0 to 1 in evenly-spaced intervals of 0.1. Model fitting in this way leads to a set of model coefficients $\hat{\boldsymbol{\theta}}$ for a particular set of predictors (i.e., genome-wide per-gene methylation measures), with one

coefficient per predictor, defining those predictors which are present in the model (i.e., predictors with corresponding non-zero coefficients), and their relative weightings.

The fitted model coefficients $\hat{\boldsymbol{\theta}}$ calculated according to Eqs (1) and (2) and the training set data were used to calculate a score $\hat{\boldsymbol{\theta}}^{\top}\mathbf{x}_j$ for each patient j, based on the corresponding per-gene methylation measures $\mathbf{x}_j$. These scores were then used to divide the training set into tertiles, defining high and low risk groups. The cutoffs defining the top and bottom tertiles in the training set were then used to divide the test set into three portions, and those most and least at risk (i.e., those test set patients with scores above the top cutoff, and below the bottom cutoff) were compared by Mantzel-Haenszel test, stratified for age, stage and residual disease (disease grade was not associated with survival for this data set), to assess the ability of this model fit to blindly predict patient survival, adjusted for significant clinical covariates. Upper and lower tertiles were compared here as previously by other authors [9] for the OC prognostic signature generation, and the reasoning for doing so in this discovery stage, rather than comparing two groups separated by the median score, was in order to prioritise larger effect sizes. If the samples were split into two groups divided by the median score, relatively small differences in the per-gene methylation measures used to generate this score might result in patients being categorised as high or low risk, with corresponding significant test results from this small variation between patients. Comparison of upper and lower tertiles would be expected to be more robust / stable with respect to such small differences in per-gene methylation measures.

Due to the heterogeneity in the main OC data set which was used to generate the OC prognostic signature, each randomly-selected training set which the Elastic Net model was fitted to lead to a different set of genes being chosen. In order to infer a consistent OC prognostic signature from this data set, i.e., a consensus, the same process of randomly partitioning the data and fitting the model was repeated a total of 10^5 times for the gene-body IGV measure. Of these, 8281 model fits were able to significantly predict survival in the respective test set (FDR $q < 0.1$). To generate the OC prognostic signature, genes were first ranked by how many of these 8281 significant model fits they appeared in. In the case of ties, genes were additionally ranked by, for each model fit, calculating the proportion of the sum of the absolute coefficient values for that model, which each gene selected as part of that model accounted for, and then comparing, for each tied gene, the mean of these proportions for that gene, across all the models it was selected as part of. Genes were assigned significance according to how many models they were selected as being part of, y, out of the total $k = 8281$ models selected as significantly associated with survival, under the null hypothesis that they were present in these observed y significant model fits by chance. If there were the same number of genes selected as part of each of these 8281 model fits, then this significance under the null hypothesis might be modelled by a binomial distribution, with the probability p_b of any gene being selected by chance as part of one model fit approximated by $p_b = f/m$, where f is the number of genes selected as part of each and every model fit, and m is the total number of genes for which gene-body methylation variance information is available. The probability of seeing a gene purely by chance in at least y model fits, out of a possible total of k, with constant probability p_b of appearing in each of these k models, would then be given by Eq (3),

$$P(Y \geq y) = \sum_{r=y}^{k} \left[\binom{k}{r} p_b^r (1 - p_b)^{(k-r)} \right]. \tag{3}$$

However, the number of genes selected, f, as part of each model, varies considerably (from 7 to 1697), and consequently p_b cannot be assumed to be constant. Alternatively, p_b could be modelled as being variable and bounded on [0, 1], with a corresponding probability distribution π_b

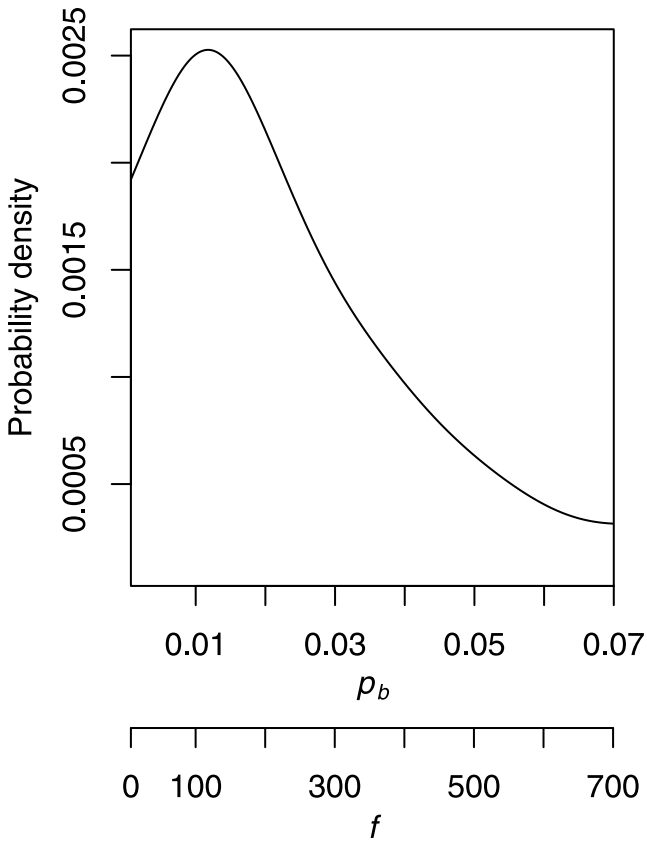


Fig 7. Probability density distribution of the probabilities of a gene being included in a fitted model. The plot shows a kernel-smoothed empirical estimate of the probability density distribution of the number of genes included in each model, f, over the 8281 significant gene body methylation variance model fits, with corresponding probability of a gene being included in a model $p_b = f/m$, where m is the number of genes with gene body methylation variance information available.

doi:10.1371/journal.pone.0143178.g007

(p_b). The distribution $\pi_b(p_b)$ can be estimated as the observed distribution of f among the $k = 8281$ significant model fits, again using $p_b = f/m$. This leads to a modelled probability, Eq (4), of seeing any gene at least y times out of k model fits purely by chance, with p_b variable and with its distribution $\pi_b(p_b)$ empirically estimated as $\hat{\pi}_b(p_b)$,

$$P(Y \geq y) = \sum_{r=y}^{k} \left\{ \int_0^1 \hat{\pi}_b(p_b) \left[\binom{k}{r} p_b^r (1-p_b)^{(k-r)} \right] \mathrm{d}p_b \right\}, \quad (4)$$

with the square brackets included in Eq (4) to highlight the comparison with Eq (3). In practice, the integral in Eq (4) is replaced with a sum over the observed values of p_b, as calculated from the observed values of f, which range between 7 and 1697. A kernel-smoothed plot of $\hat{\pi}_b(p_b)$, the empirical probability density distribution of f and corresponding p_b, appears in Fig 7.

Calculation of the DNAm IGV ovarian cancer prognostic score

Clustering was performed to identify groups of genes in the OC prognostic signature with similar patterns of IGV across patients. The clustering was carried out separately for genes individually associated with worse patient survival outcome for increased IGV ('hyper' genes) and for decreased IGV ('hypo' genes). Consensus clustering [19] was used for the clustering, with a

hierarchical clustering inner loop, using $1 - \rho$ as the distance measure, where ρ is the Spearman rank correlation coefficient. The following additional settings were used: probability of selecting a sample = 0.8, probability of selecting a feature = 1, number of resamplings = 10^5, maximum number of clusters = 20.

The discovered clusters were then filtered (to remove noise, and uncertainty associated with trends inferred from small groups of genes in these genome-wide data), retaining only those clusters which contained at least 10 genes, and only those clusters with mean IGV significantly associated with patient survival outcome. After filtering, four clusters remained, for two of which an increase in the cluster mean IGV was associated with worse patient survival outcome (called 'hyper 1' and 'hyper 2'), and for two of which a decrease in the cluster mean IGV was associated with worse survival outcome (called 'hypo 1' and 'hypo 2'). The IGV cluster scores were then calculated, as the means of the IGV of the genes each of these four clusters.

In order to calculate the IGV prognostic score from these components, a Cox model (adjusted for clinical covariates) was fitted to these four IGV cluster scores. The coefficients for this model (standardised by the variance of the predictors) are fairly similar for each of the clusters (hyper 1: 0.22; hyper 2: 0.25; hypo 1: 0.23; hypo 2: 0.30), indicating that each cluster is important to the model, and to the prognostic predictions. The median of the IGV prognostic score calculated from this Cox model was used to divide the 218 patients in the main DNAm OC data-set used to derive the OC prognostic signature, into better and worse prognostic groups.

Validation of the ovarian cancer prognostic signature

The DNAm prognostic signature derived from the OC data set was validated in two independent DNAm data sets. The first of these data sets was taken from another study of OC (n = 198), and was supplied by the Mayo Clinic. The second of these data sets was taken from a study of uterine corpus endometrioid carcinoma (UCEC) (n = 358), and was downloaded from *The Cancer Genome Atlas* (TCGA) project [21].

The IGV prognostic score was similarly calculated by fitting a Cox model to the four IGV cluster scores in the main OC DNAm data set, adjusted for clinical covariates, then applying this model to the equivalent IGV cluster scores in the Mayo Clinic OC and the TCGA UCEC validation sets. In order to make prognostic predictions in these independent data sets using only the DNAm data, the model was used to calculate the IGV prognostic score for the samples in the independent data sets from the fitted model coefficients corresponding to IGV cluster scores only, and not the clinical covariates. This IGV prognostic score was used to define better and worse prognostic groups in the independent data sets, separated by the median IGV prognostic score in the main OC data set. These prognostic groups were then compared, assessing statistical significance with univariate and multivariate Cox regression (i.e., respectively without and with adjustment for the clinical covariates).

Comparison of IGV with Intra-Tumour Heterogeneity

Intra-tumour methylation heterogeneity was assessed in terms of cross-sample variability of methylation, where the samples are taken from the same patient. The resulting patterns and relationships are compared with cross-sample variability of methylation in the main OC data set, where the samples are now from different patients. Cross-sample variability of methylation is estimated by first calculating the variance of the methylation measurements across all samples for a particular probe, and then taking the mean of these probe variances for all the probes in a gene (gene body probes only). This mean cross-sample methylation variance was compared to the mean IGV of the same gene, which is calculated by taking the mean of the IGV for that gene across the same samples as were used to calculate the cross-sample methylation

variance. Cross-sample methylation variance was then analysed as a function of IGV by estimating $E(y|x)$, where y represents cross-sample methylation variance and x represents IGV, by fitting spline curves. This resulted in one best-fit curve per patient for the EC and prostate cancer intra-tumour heterogeneity datasets, and one best-fit curve for all the patients for each of the main OC and TCGA BRCA basal datasets.

Testing Transcription-factor binding correlation with IGV

We examined transcription factor binding to the OC prognostic signature genes, using the ENCODE (Encyclopedia of DNA Elements) chromatin immunoprecipitation (ChIP) data [24], with the ANNOVAR software [48]. Transcription factor binding site information was available, for the gene body regions defined, for 55 transcription factors. Each of these TFs was tested for significant over or under enrichment binding to the genes of each of the four prognostic signature clusters, with Fisher's exact test. We also tested the correlation of the expression level of each of these 55 TFs, with the IGV of genes the TF binds to, and the genes the TF does not bind to. We used a Kolmogorov-Smirnov test to assess whether, for each TF, there is significantly more positive, or more negative, correlation with IGV of the genes it binds to, compared to genes it does not. For this expression correlation analysis, we used the 42 TCGA BRCA basal samples with both expression and DNAm data available, because it was comprehensively demonstrated by the TCGA consortium that high-grade serous ovarian and uterine and BRCA basal cancers are extremely molecularly similar [25], and we also established a high degree of similarity of behaviour between our prognostic signature genes in OC, and these TCGA BRCA basal samples.

Ethics Statement

The use of tumour tissue has been approved by the local ethical committees of the contributing institutions: Studying the samples contributed from Rotterdam has been approved by the local Rotterdam Medical Ethics Committee (MEC-2008-183), performed in accordance with the Code of Conduct of the Federation of Medical Scientific Societies in the Netherlands (http://www.fmwv.nl). The Regional Committee for Medical Research Ethics in Norway approved the study (for ovarian cancer patients diagnosed in Oslo before 2007, exemption from obtaining informed consent was received as the majority of ovarian cancer patients were dead at the time the application was evaluated; patients diagnosed after 2007 signed general consent allowing for use of the tumours for research purposes). Written informed consent for the use of tumour tissue and prospective clinical data collection was obtained from all patients and approved by the Leuven ethics committee. The use of cancer samples from Bergen was approved by the Regional Research Ethics Committee in Medicine and patients have given their written informed consent to use their sample for research. Patients whose samples were used from the Mayo Clinic gave informed consent and the Mayo Clinic Institutional Review Board approved the study. No identifying patient information was available to us. The data have not been published before.

Supporting Information

S1 Fig. Mean IGV across patients, for the genes of each cluster.
(TIF)

S2 Fig. Mean gene-body methylation level, across patients, for the genes of each cluster.
(PDF)

S1 File. Supplementary Tables.
(PDF)

Acknowledgments

We are very grateful to all specimen donors and research groups involved in providing the data used in this study via TCGA and GEO. This work was funded (MW and AJ) by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 305428 (Project EpiFemCare), by the National Institute for Health Research University College London Hospitals Biomedical Research Centre, and by the Eve Appeal and the European Network Translational Research in Gynaecological Oncology (ENTRIGO) of the European Society of Gynaecological Oncology (ESGO). TEB received funding from the UK Engineering and Physical Sciences Research Council (ESPRC) and the UK Medical Research Council (MRC) via UCL CoMPLEX. ELG received funding from the Fred C. and Katherine B. Andersen Foundation, NIH grants R01-CA122443, P50-CA136393 (the Mayo Clinic Ovarian Cancer SPORE) and P30-CA15083. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived and designed the experiments: TEB MW. Performed the experiments: AJ JMC. Analyzed the data: TEB MW. Contributed reagents/materials/analysis tools: ELG BLF EW HBS BD CGT SL IV EMJJB. Wrote the paper: TEB MW.

References

1. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nature Reviews Genetics. 2006; 7(1):21–33. doi: 10.1038/nrg1748 PMID: 16369569
2. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nature Reviews Genetics. 2002; 3(6):415–428. PMID: 12042769
3. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: A Cancer Journal for Clinicians. 2011; 61(2):69–90.
4. Greenlee RT, Hill-Harmon MB, Murray T, Thun M. Cancer statistics, 2001. CA: A Cancer Journal for Clinicians. 2001; 51(1):15–36.
5. Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, et al. Epigenetic stem cell signature in cancer. Nature Genetics. 2006; 39(2):157–158. doi: 10.1038/ng1941 PMID: 17200673
6. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. Nature Reviews Genetics. 2007; 8(4):253–262. doi: 10.1038/nrg2045 PMID: 17363974
7. Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. Nature Reviews Genetics. 2012; 13(2):97–109. PMID: 22215131
8. Bartlett TE, Olhede SC, Zaikin A. A DNA Methylation Network Interaction Measure, and Detection of Network Oncomarkers. PloS One. 2014; 9(1):e84573. doi: 10.1371/journal.pone.0084573 PMID: 24400102
9. Zhuang J, Jones A, Lee SH, Ng E, Fiegl H, Zikan M, et al. The Dynamics and Prognostic Potential of DNA Methylation Changes at Stem Cell Gene Loci in Women's Cancer. PLoS Genetics. 2012; 8(2): e1002517. doi: 10.1371/journal.pgen.1002517 PMID: 22346766
10. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. Significance analysis and statistical dissection of variably methylated regions. Biostatistics. 2012; 13(1):166–178. doi: 10.1093/biostatistics/kxr013 PMID: 21685414
11. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, et al. Increased methylation variation in epigenetic domains across cancer types. Nature Genetics. 2011; 43(8):768–775. doi: 10.1038/ng.865 PMID: 21706001
12. Bartlett TE, Zaikin A, Olhede SC, West J, Teschendorff AE, Widschwendter M. Corruption of the Intra-Gene DNA Methylation Architecture Is a Hallmark of Cancer. PloS One. 2013; 8(7):e68285. doi: 10.1371/journal.pone.0068285 PMID: 23874574

13. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–320. doi: 10.1111/j.1467-9868.2005.00503.x
14. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2011; 39(5):1–13. doi: 10.18637/jss.v039.i05
15. Jojic V, Shay T, Sylvia K, Zuk O, Sun X, Kang J, et al. Identification of transcriptional regulators in the mouse immune system. Nature Immunology. 2013; 14(6):633–643. doi: 10.1038/ni.2587 PMID: 23624555
16. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57(1):289–300.
17. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286 (5439):531–537. doi: 10.1126/science.286.5439.531 PMID: 10521349
18. Valk PJ, Verhaak RG, Beijen MA, Erpelinck CA, van Doorn-Khosrovani SBvW, Boer JM, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. New England Journal of Medicine. 2004; 350(16):1617–1628. doi: 10.1056/NEJMoa040465 PMID: 15084694
19. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning. 2003; 52(1–2):91–118. doi: 10.1023/A:1023949509487
20. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. American Journal of Epidemiology. 2007; 165(6):710–718. doi: 10.1093/aje/kwk052 PMID: 17182981
21. Collins F, Barker A. Mapping the cancer genome. Scientific American Magazine. 2007; 296(3):50–57. doi: 10.1038/scientificamerican0307-50
22. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014; 346(6206):251–256. doi: 10.1126/science.1253462 PMID: 25301630
23. Brocks D, Assenov Y, Minner S, Bogatyrova O, Simon R, Koop C, et al. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. Cell reports. 2014; 8(3):798–806. doi: 10.1016/j.celrep.2014.06.053 PMID: 25066126
24. Consortium EP, et al. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004; 306 (5696):636–640. doi: 10.1126/science.1105136
25. Network CGA, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490 (7418):61–70. doi: 10.1038/nature11412
26. Nitzsche A, Paszkowski-Rogacz M, Matarese F, Janssen-Megens EM, Hubner NC, Schulz H, et al. RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. PloS One. 2011; 6(5):e19470. doi: 10.1371/journal.pone.0019470 PMID: 21589869
27. Attanasio C, Nord AS, Zhu Y, Blow MJ, Biddie SC, Mendenhall EM, et al. Tissue-specific SMARCA4 binding at active and repressed regulatory elements during embryogenesis. Genome research. 2014; 24(6):920–929. doi: 10.1101/gr.168930.113 PMID: 24752179
28. Witkowski L, Carrot-Zhang J, Albrecht S, Fahiminiya S, Hamel N, Tomiak E, et al. Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type. Nature Genetics. 2014; 46(5):438–443. doi: 10.1038/ng.2931 PMID: 24658002
29. Ramos P, Karnezis AN, Craig DW, Sekulic A, Russell ML, Hendricks WP, et al. Small cell carcinoma of the ovary, hypercalcemic type, displays frequent inactivating germline and somatic mutations in SMARCA4. Nature genetics. 2014; 46(5):427–429. doi: 10.1038/ng.2928 PMID: 24658001
30. Jelinic P, Mueller JJ, Olvera N, Dao F, Scott SN, Shah R, et al. Recurrent SMARCA4 mutations in small cell carcinoma of the ovary. Nature Genetics. 2014; 46(5):424–426. doi: 10.1038/ng.2922 PMID: 24658004
31. Gartel A. A new mode of transcriptional repression by c-Myc: methylation. Oncogene. 2006; 25 (14):1989–1990. doi: 10.1038/sj.onc.1209101 PMID: 16170342
32. Yokomori N, Tawata M, Saito T, Shimura H, Onaya T. Regulation of the rat thyrotropin receptor gene by the methylation-sensitive transcription factor GA-binding protein. Molecular Endocrinology. 1998; 12 (8):1241–1249. doi: 10.1210/mend.12.8.0142 PMID: 9717849
33. Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PA, Rappsilber J, et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature. 2011; 473(7347):343–348. doi: 10.1038/nature10066 PMID: 21490601

34. Kirillov A, Kistler B, Mostoslavsky R, Cedar H, Wirth T, Bergman Y. A role for nuclear NF–κB in B–cell–specific demethylation of the Igκ locus. Nature Genetics. 1996; 13(4):435–441. doi: 10.1038/ng0895-435 PMID: 8696338

35. Zechner U, Seifert D, Schneider E, El Hajj N, Navarro B, Kondova I, et al. Different DNA methylation of FOXP2 target genes in adult cortices of humans and chimpanzees. In: Proceedings of the Annual Meeting of the American Society of Human Genetics. American Society of Human Genetics; 2012. p. 3266W.

36. Coulson JM. Transcriptional regulation: cancer, neurons and the REST. Current biology. 2005; 15(17): R665–R668. doi: 10.1016/j.cub.2005.08.032 PMID: 16139198

37. Gustems M, Woellmer A, Rothbauer U, Eck SH, Wieland T, Lutter D, et al. c-Jun/c-Fos heterodimers regulate cellular genes via a newly identified class of methylated DNA sequence motifs. Nucleic Acids Research. 2014; 42(5):3059–3072. doi: 10.1093/nar/gkt1323 PMID: 24371273

38. Ng CW, Yildirim F, Yap YS, Dalin S, Matthews BJ, Velez PJ, et al. Extensive changes in DNA methylation are associated with expression of mutant huntingtin. Proceedings of the National Academy of Sciences. 2013; 110(6):2354–2359. doi: 10.1073/pnas.1221292110

39. Malone CS, Miner MD, Doerr JR, Jackson JP, Jacobsen SE, Wall R, et al. CmC (A/T) GG DNA methylation in mature B cell lymphoma gene silencing. Proceedings of the National Academy of Sciences. 2001; 98(18):10404–10409. doi: 10.1073/pnas.181206898

40. Zhu WG, Srinivasan K, Dai Z, Duan W, Druhan LJ, Ding H, et al. Methylation of adjacent CpG sites affects Sp1/Sp3 binding and activity in the p21Cip1 promoter. Molecular and Cellular Biology. 2003; 23 (12):4056–4065. doi: 10.1128/MCB.23.12.4056-4065.2003 PMID: 12773551

41. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics. 2012; 13(7):484–492. doi: 10.1038/nrg3230 PMID: 22641018

42. Consortium EP, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. doi: 10.1038/nature11247

43. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer cell. 2014; 26(4):577–590. doi: 10.1016/j.ccr.2014.07.028 PMID: 25263941

44. Schwarz RF, Ng CK, Cooke SL, Newman S, Temple J, Piskorz AM, et al. Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. PLoS medicine. 2015; 12(2): e1001789. doi: 10.1371/journal.pmed.1001789 PMID: 25710373

45. Mroz EA, Tward AM, Hammon RJ, Ren Y, Rocco JW. Intra-tumor Genetic Heterogeneity and Mortality in Head and Neck Cancer: Analysis of Data from The Cancer Genome Atlas. PLoS medicine. 2015; 12 (2):e1001786. doi: 10.1371/journal.pmed.1001786 PMID: 25668320

46. Beck AH. Open Access to Large Scale Datasets Is Needed to Translate Knowledge of Cancer Heterogeneity into Better Patient Outcomes. PLoS medicine. 2015; 12(2):e1001794. doi: 10.1371/journal.pmed.1001794 PMID: 25710538

47. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002; 30(1):207–210. doi: 10.1093/nar/30.1.207 PMID: 11752295

48. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research. 2010; 38(16):e164–e164. doi: 10.1093/nar/gkq603 PMID: 20601685