SUPPLEMENTARY FILE 1: PSEUDO-CODE FOR THE A-CLUSTERING AND d_{bp}-MERGE ALGORITHMS

Algorithm 1 (*A-clustering*) Given ordered methylation measures, by chromosomal location, at m sites for n subjects, an m-vector of chromosomal locations L, corresponding m-vector of cluster assignments, distance measure (e.g. Spearman correlation), distance measure threshold $\bar{\mathcal{D}}$, base pair distance $\bar{d}_{bp}$ and *type* (*single, average* or *complete*)

```
(1) Set the cluster indicator $ind\_current\_cluster \leftarrow 1$.
      • The matching set $current\_cluster$ contains all the indices of sites belonging to
        cluster indicated by $ind\_current\_cluster$.
      • The set $next\_cluster$ contains all the indices of sites belonging to cluster
        indicated by $ind\_current\_cluster + 1$.
(2) Until no more clusters can be merged:
    (a) Let merge be true if $dist(current\_cluster, next\_cluster) \leq \bar{\mathcal{D}}$ (depend on type),
        and if also $\min(L(next\_cluster)) - \max(L(current\_cluster)) < \bar{d}_{bp}$.
    (b) Repeat while merge is True: $current\_cluster \leftarrow current\_cluster \cup next\_cluster$
        by updating cluster assignments
    (c) If merge is False: $ind\_current\_cluster \leftarrow (ind\_current\_cluster+1)\ \%\ current\_number\_of\_clusters$
        (modulo)
(3) Output cluster assignments.
```

Note that we know that “no more clusters can be merged” after cycling through all clusters and merge none.

Algorithm 2 (d_{bp}*-merge*) Given ordered methylation measures, by chromosomal location, at m sites for n subjects, an m-vector of chromosomal locations L, distance measure (e.g. Spearman correlation), a threshold $\bar{\mathcal{D}}$, and base-pair distance $\underline{d}_{bp}$. Let $cluster(site_k)$ be the cluster of site k; initially it is the kth site alone.

```
(1) Set the site indicator $ind\_current\_site \leftarrow 1$. The matching site is $current\_site$
(2) While $ind\_current\_site < number\_of\_sites$:
    (a) Let $\mathrm{site}_1, \ldots, \mathrm{site}_k$ be all sites such that $L(\mathrm{site}_l) - L(current\_site) \leq \underline{d}_{bp}$
        ordered by their distance from $current\_site$.
    (b) Set $merge \leftarrow$ False, $l \leftarrow k$.
    (c) While merge is False and $l > ind\_current\_site$:
          • If $dist(current\_site, site_l) < \bar{\mathcal{D}}$
              – Set $cluster(current\_site, \ldots, \mathrm{site}_l) \leftarrow cluster(current\_site)$.
              – Set $merge \leftarrow$ True.
          • Else $l \leftarrow l - 1$
```

The following figure is Figure 1 from the paper, showing a difference in clustering results without/with using d_{bp}-merge initiation,

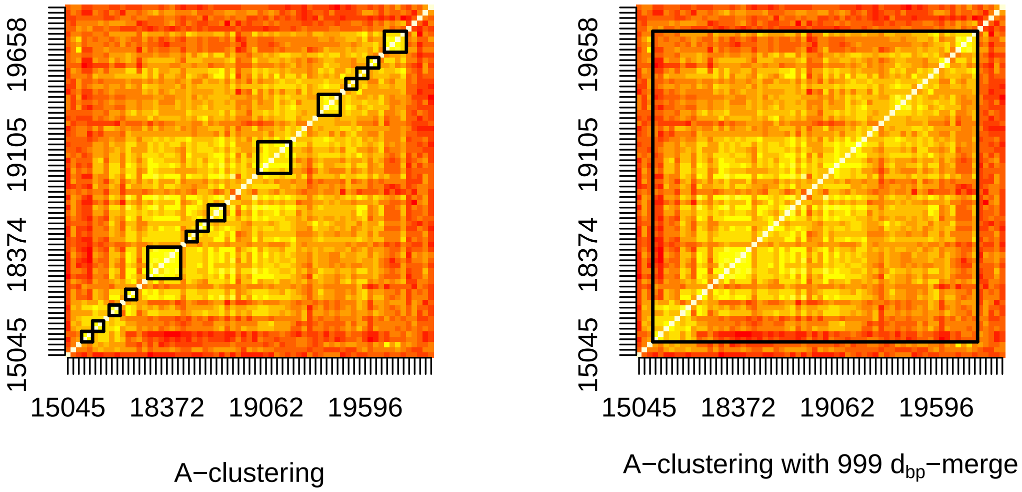

FIGURE 1. Figure 1 from the paper - in color. Left panel: d_{bp}-merge was not used. Right panel: d_{bp}-merge initiation followed by Aclust.