Robust foreground detection in somatic copy number data

Aditya Deshpande[1,5], Trent Walradt[2,5], Ya Hu[3,5], Amnon Koren[3], and Marcin Imielinski[2,4,5,6]

[1]Tri Institutional PhD Program in Computational Biology and Medicine, Graduate School of Medical Sciences, Weill Cornell Medicine, 1300 York Avenue, New York, NY 10065, USA
[2]Weill Cornell Medical College, 445 East 69th Street, New York, NY 10065, USA,
[3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA
[4]Pathology and Laboratory Medicine, Weill Cornell Medicine, 1300 York Avenue, New York, NY 10065, USA
[5]New York Genome Center, 101 Avenue of the Americas, New York, NY 10013, USA,
[6]Institute for Computational Biomedicine, Weill Cornell Medicine, 1300 York Avenue, New York, NY 10065, USA

Sensitive detection of somatic copy number alterations (SCNA) in cancer genomes is confounded by "waviness" in read depth data. We present `dryclean`, a signal processing algorithm to optimize SCNA detection in whole genome (WGS) and targeted sequencing platforms through foreground detection and background subtraction of read depth data. Application of `dryclean` to WGS demonstrates that WGS waviness is driven by replication timing. Re-analysis of thousands of tumor profiles reveals that `dryclean` provides superior detection of biologically relevant SCNAs relative to state-of-the-art algorithms. Applied to *in silico* tumor dilutions, `dryclean` improves the sensitivity of relapse detection 10-fold relative to current standards. `dryclean` is available as an R package in the GitHub repository https://github.com/mskilab/dryclean

Background

Accurate somatic copy number alteration (SCNA) detection is a crucial aspect of precision oncology and cancer genome analysis. Therapeutically actionable SCNAs include trastuzumab-sensitizing *ERBB2* amplifications in breast cancer (1) and *CDKN2A* deletions or *CDK4* amplifications identifying responders to *CDK4/6* inhibitors in glioblastoma (2). SCNAs resulting in the loss of *RB* and *TP53* are associated with later "small-cell conversion", a mechanism of aggressive erlotinib-resistant relapse in lung adenocarcinoma (3). In a modern oncology practice, the decision to provide or deny a certain therapy may be heavily influenced by sequencing results (4). Consequently, it is vital that these tests achieve an adequate level of sensitivity. Accurate SCNA detection is also crucial for enabling discovery of recurrent patterns of structural variation or genotyping SCNAs in samples where somatic clones are highly diluted against a euploid background (e.g. solid and liquid biopsy relapse monitoring, clonal hematopoiesis detection) (5–7).

A key challenge in SCNA inference from DNA sequencing experiments is noise or bias in read depth data (8). SCNA inference relies on an affine relationship between copy number (CN) and read depth to identify segments of discrete CN states. This relationship is distorted by assay– and batch-specific effects, including mappability bias, GC bias, library quality, input quantity, and heterogeneity in capture probe affinity (9, 10). Though many methods exist for SCNA inference in whole genome (WGS), whole exome (WES), or targeted panel (TP) sequencing, most take the tumor-normal ratio (TNR) as input (6, 11–16). The reason for this normalization, in theory, is to remove systematic biases since both tumor and normal samples are processed similarly (e.g. similar capture probes, library preparation, read lengths).

As a ratio of two noisy signals, however, the TNR is sensitive to subtle or stochastic differences in batch effects, including low-quality normal libraries or specimens (11). Furthermore, the presence of inherited CN events can lead to puzzling patterns in the TNR, such as foci of apparent CN gain when a region containing a focal germline deletion undergoes loss of heterozygosity. Consequently, reliance on TNR may lead to misinterpretation of data (false negatives and false positives). Finally, a matched normal may be absent for certain samples, such as a cancer cell line or archival formalin-fixed paraffin-embedded (FFPE) tissue. Here, computation of TNR from an arbitrary or composite diploid sample may provide inadequate normalization or even add distortion to the read depth signal.

An ideal read depth normalization method would flatten assay– and batch-specific biases (termed "background") while retaining transitions associated with CN changes (termed "foreground"). Previous approaches for read depth normalization include supervised algorithms that explicitly correct for known factors, specifically GC bias and mappability (10, 17)). As we show, GC and mappability correction fails to remove significant biases that plague both whole exome and whole genome sequencing read depth data, causing artifacts in downstream SCNA analyses. To overcome the significant limitations of these state-of-the-art pipelines, we have developed `dryclean`. `dryclean` learns a background model of biological and technical bias through the analysis of a panel of diploid normal samples (PON) using a signal processing method previously used for object detection in video data (18). Applying this model to tumor data, `dryclean` enriches for SCNA signal while subtracting the assay– and batch-specific background. Through reanalysis of large sequencing datasets from the Cancer Cell Line Encyclopedia (CCLE) and the Cancer Genome Atlas (TCGA), we reveal footprints of DNA replication in WGS background, improve the calling of biologically significant and clinically actionable SCNA, and demonstrate sensitive detection of SCNA in very dilute ($<1\%$) tumor samples (such

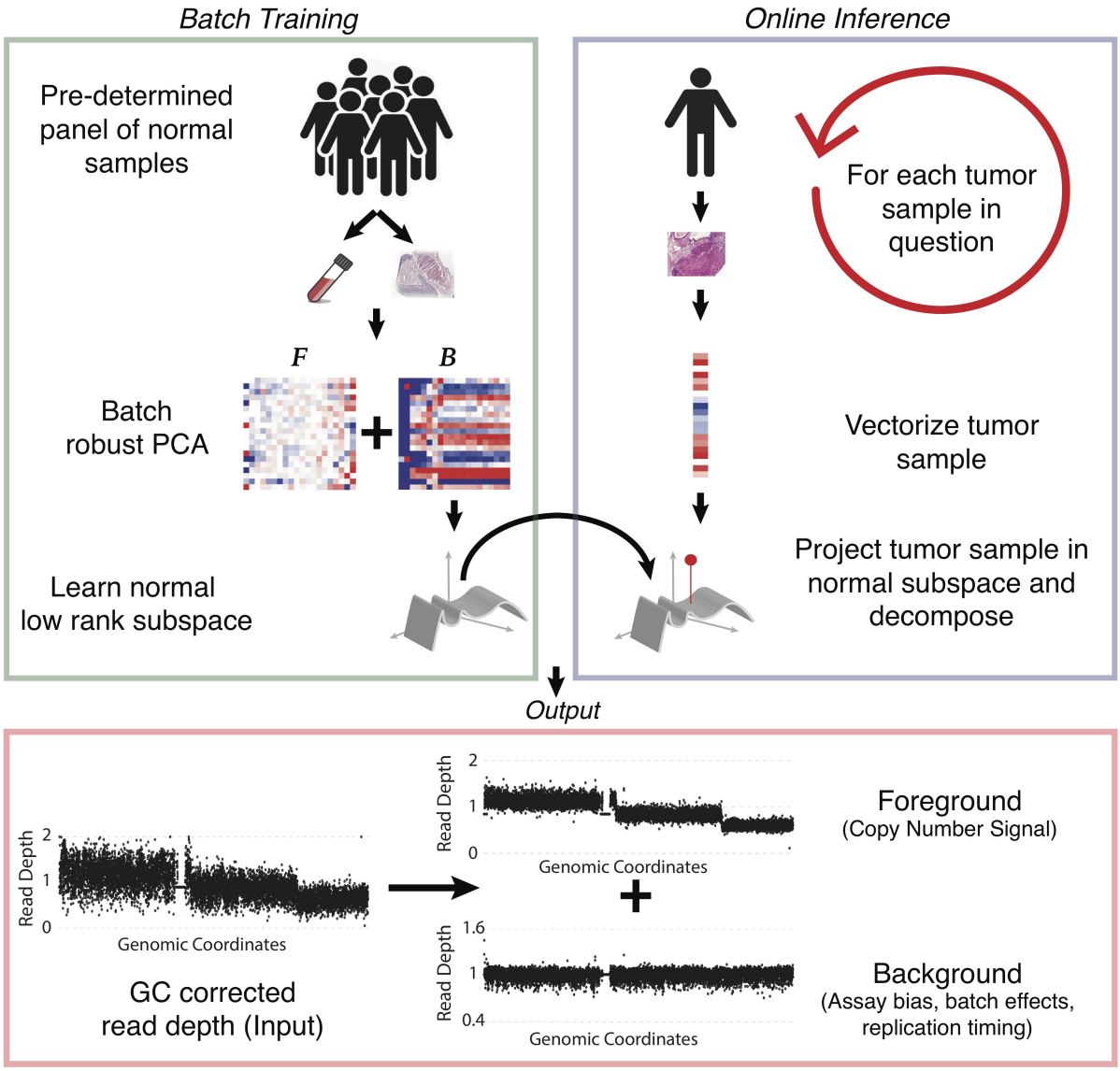


Fig. 1. `dryclean` schematic The `dryclean` pipeline separates foreground (SCNA signal) from background (technical and biological bias) in read depth data by adapting an algorithm used to detect objects in video surveillance data. `dryclean` learns a low rank subspace across a panel of normal (PON) samples using rPCA and applies this to a tumor sample to infer SCNA. The three boxes indicate stages of the `dryclean` pipeline. First, (green box) a PON is created from log transformed, zero centered, GC and mappability corrected read depths across a set of normal samples and genomic windows. Batch rPCA decomposition yields a low rank $\mathbf{B}$ matrix, whose columns span a subspace of biases that define the assay "background". A second step (purple box) projects a tumor onto the background subspace to identify two vectors: the low-rank projection ($\mathbf{b}$) representing background bias and sparse residual ($\mathbf{f}$) corresponding to foreground or SCNA (red box).

as those that might be obtained during liquid or solid biopsy relapse monitoring).

Results

Robust foreground detection in read depth data. A central premise of `dryclean` is that genome-wide read depth data is shaped by an assay, batch, and sample specific background that arises from the combination of a small set of (unknown) factors, each affecting a large fraction of samples. In contrast, each tumor sample harbors a unique SCNA profile, comprising both broad (e.g. arm level gains and losses) and focal (e.g. double minute amplifications) variants that are rarely seen in normal samples.

To apply this intuition to cancer genomic read depth data, we borrow a powerful tool from computer vision and signal processing called robust principal components analysis (rPCA). A key previous application of rPCA is object detection in video surveillance (18). Here, a set of video images comprise a slowly changing background occasionally perturbed by a fast moving object, which defines the foreground of a subset of pixels in a handful of frames. Through the solution of an optimization, rPCA can decompose the matrix of n video images, each with m pixels, as the sum of a low rank (slowly changing) background matrix and sparse foreground matrix.

Similarly, we apply rPCA to detect cancer SCNA as a sparse and high amplitude perturbation against a relatively constant background defined by $n-1$ normal diploid sam-

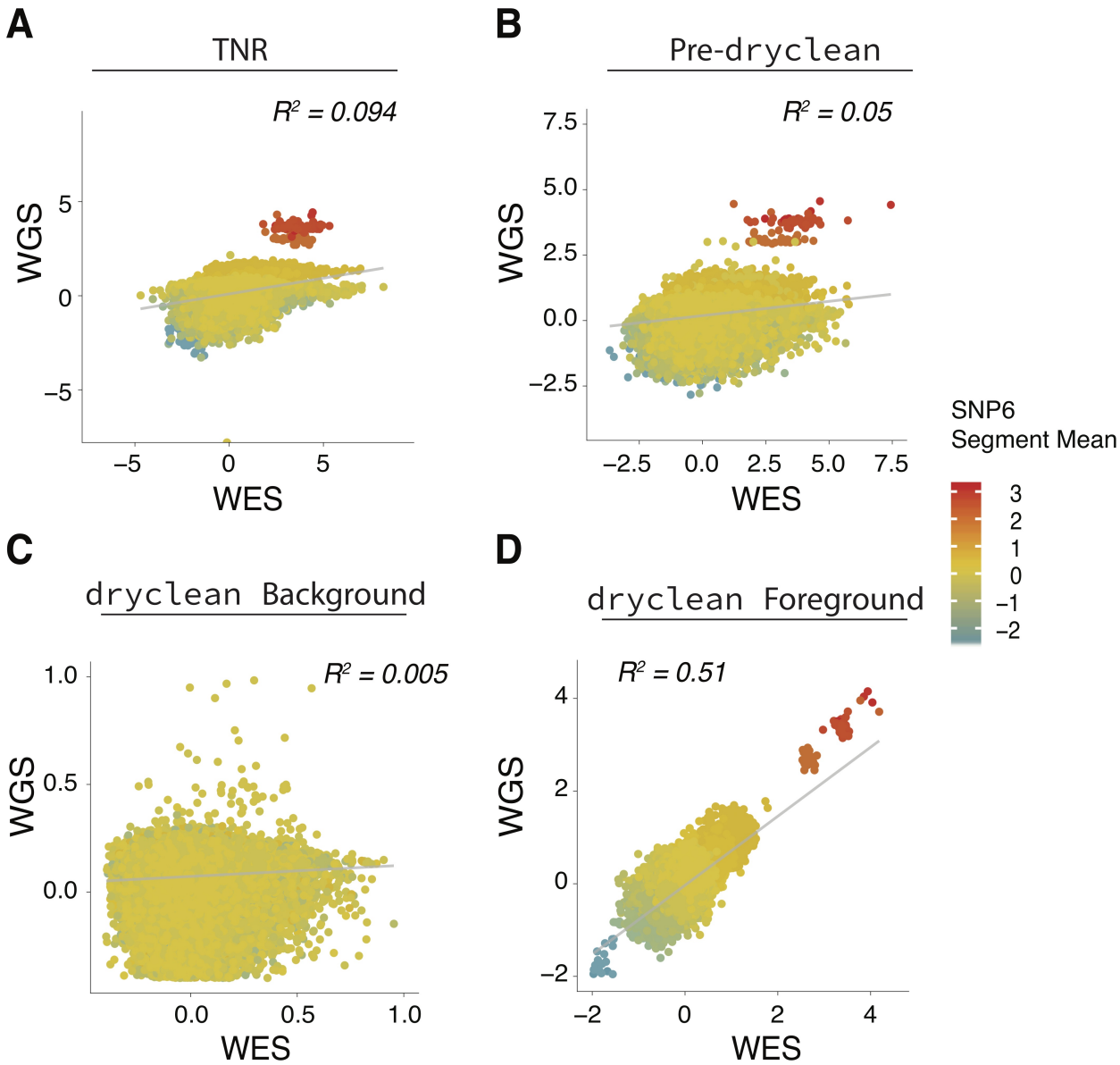


Fig. 2. `dryclean` foreground reveals SCNA signal across diverse sequencing assays For (A-D), WES and WGS read depth data are shown for a TCGA tumor and normal pair (TCGA-02-2483). Each point on the scatter plot represent GC and mappability corrected (17) read count for bins present in both WGS and WES. The color of each point represents the copy number signal from segmented SNP6 gold standard data (log). (A) The TNR shows poor inter-platform correlation (R^2 = 0.094) between WES and WGS, which suffer from different assay and batch specific biases. The copy number separation is poor and the low copy number regions are indistinguishable. (B) This correlation is similarly low even when we compare the pre-`dryclean`ed vectors representing tumor read depth following GC and mappability correction (R^2 = 0.05). As with TNR, the copy states are weakly separated. (C) The `dryclean` inferred background for WES and WGS data (decomposing $\mathbf{m}$ data against their respective PON) represents assay and batch specific biases, which (as expected) demonstrates little correlation (R^2 = 0.005) between platforms. Regions from all copy state are intermixed as we would expect from the background. (D) The `dryclean` inferred foreground shows improved correlation between WGS and WES. (R^2 = 0.51) and captures discrete clusters of read depth representing SCNAs. Notable are the clearly distinct copy states that track well with SNP6 data.

ples, each comprising read depth data across a set of m genomic bins. Formally, we decompose the input matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ of read depth data across $n-1$ normal samples and one tumor sample into a low-rank matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$ and sparse matrix $\mathbf{F} \in \mathbb{R}^{m \times n}$ through the solution of a constrained optimization:

$$\begin{aligned} \underset{\mathbf{B},\mathbf{F}}{\text{minimize}} \quad & \|\mathbf{B}\|_* + \lambda \|\mathbf{F}\|_1 \\ \text{subject to} \quad & \mathbf{M} = \mathbf{B} + \mathbf{F} \end{aligned} \tag{1}$$

where $\lambda > 0$ is a regularization parameter. For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\|\mathbf{A}\|_* = \sum_i \sigma_i$ is the nuclear norm and $\|\mathbf{A}\|_1 = \sum_{ij} a_{ij}$ is the L_1 norm of a vectorized $\mathbf{A}$ matrix. Here, σ_i is the ith singular value of matrix $\mathbf{A}$ and a_{ij} is the ij component of matrix $\mathbf{A}$. In practice, we separate the solution of Eq. 1 into a computationally intensive "training" step and faster "inference" step (Figure 1, see Methods). The training step directly solves Eq. 2 for a matrix $\hat{\mathbf{M}} \in \mathbb{R}^{m \times n-1}$ of $n-1$ diploid samples to obtain $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{F}}$. The inference step computes the projection $\mathbf{b}$ of a new tumor sample $\mathbf{t} \in \mathbb{R}^m$ on the low-rank "background" column space of $\tilde{\mathbf{B}}$ and a sparse residual vector $\mathbf{f}$ that we call "foreground" (Figure 1). As we

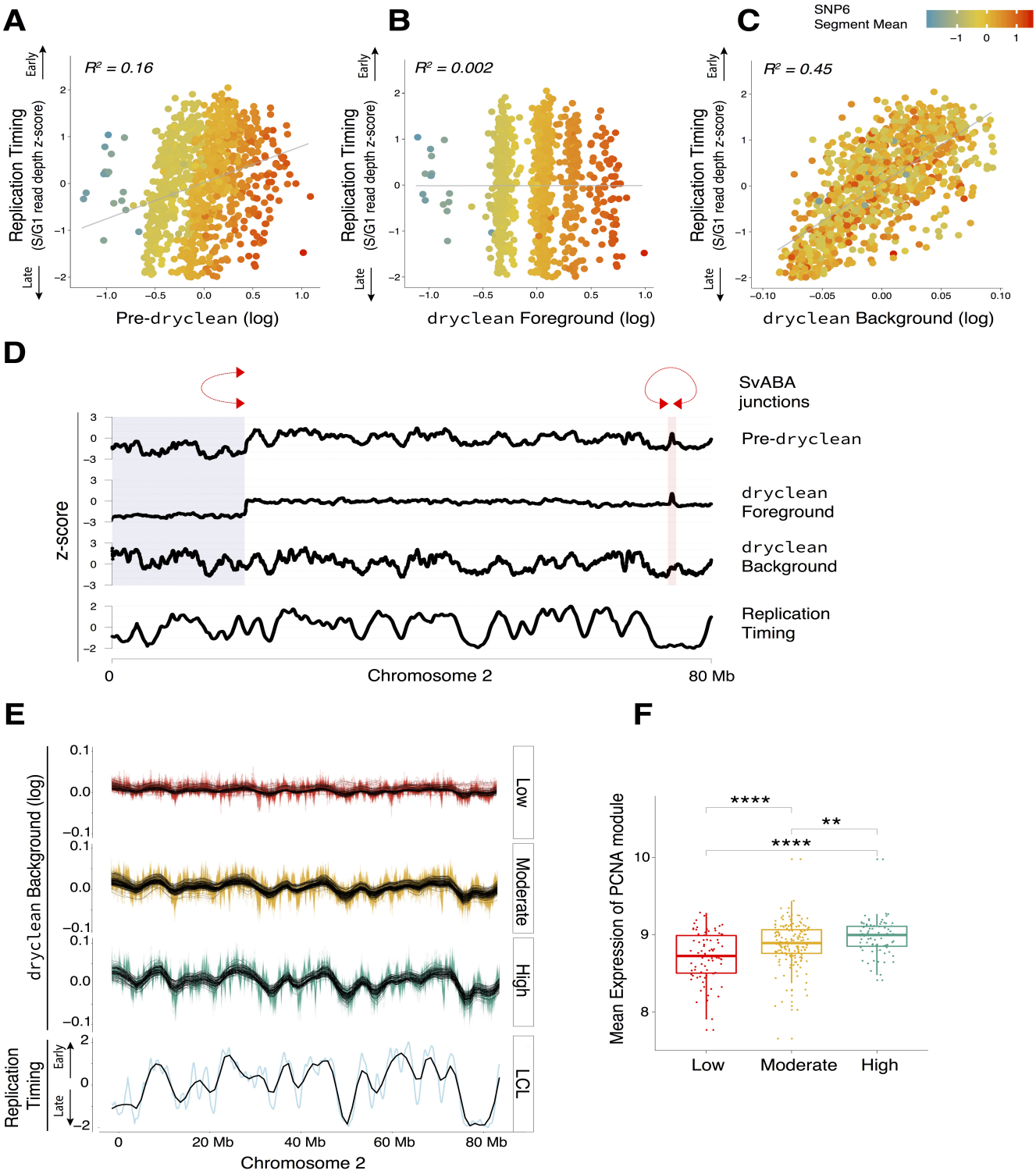


Fig. 3. `dryclean` background reveals replication waves in cancer WGS. (A) Correlation of GC and mappability corrected log read counts i.e. pre-`dryclean` data at 1000 randomly sampled loci with LCL replication timing for a CCLE sample (G16478.KE-39, stomach cancer cell line) with high replication score (see text). Similar to Figure 2, the copy number signal from SNP6 data shows poor separation. (B) `dryclean` inferred foreground ($\mathbf{f}$ vector) correlates poorly with replication timing from LCL (R^2 = 0.002), while show clear copy number states that track with SNP6. (C) `dryclean` inferred background ($\mathbf{b}$ vector) shows increased correlation (R^2 = 0.45) with replication timing profile and poor separation of copy states. (D) Snapshot of these tracks across chromosome 2p demonstrates increased correlation of `dryclean` background with replication timing relative to input i.e Pre-`dryclean`, while `dryclean` foreground signal reveals SCNA's, whose change points coincide with the locations of rearrangement junctions (red links above track). The blue shaded region highlights a deletion and red region highlights a focal amplification whose endpoints coincide with nearby concordant rearrangement junctions. (E) `dryclean`-inferred background vectors from 320 CCLE lines, grouped by quantiles into three tranches of low, medium, and high replication timing signal (see methods). Each colored track represent raw data. Black curves represent loess-smoothed fit for each sample. (F) Box plot of Proliferating Cell Nuclear Antigen (PCNA) gene expression module (described in text) demonstrates transcriptional programs associated with increased DNA synthesis in cell lines with high or medium vs low replication scores.

show below, foreground vectors represent SCNA and background vectors represent assay– and batch-specific technical and biological biases.

`dryclean` foreground enriches SCNA signal across sequencing platforms. To gain intuition into the decomposition performed by `dryclean`, we compared unsegmented WES and WGS read depth data for a TCGA tumor and normal pair (TCGA-02-2483). For this analysis, we compared the GC and mappability corrected read counts (17) for bins present in both WGS and WES. In addition, we superimposed copy number signal from the segmented SNP6 gold standard SCNA data on to these bins. Comparison of WES and WGS TNR demonstrated poor inter-platform correlation (R^2 = 0.094, Figure 2A) and poor visual evidence of discrete CN states in either dataset. This was further corroborated by a weak separation of the regions with different segmented CN states (as ascertained by SNP6). This correlation and CN separation was similarly poor when we examined the tumor-only GC and mappability corrected read counts, which comprise the numerator to the TNR and the input to `dryclean` (i.e. the $\mathbf{m}$ vector) (R^2 = 0.05, 2B).

We posited that `dryclean`-inferred background vectors for WES and WGS data represents distinct assay– and batch-specific biases. Confirming our hypothesis, the $\mathbf{b}$ vectors demonstrated little correlation (R^2 = 0.005) between the two platforms and were intermixed with respect to copy state (Figure 2C). In contrast, the foreground `dryclean` solution showed improved correlation between WGS and WES (R^2 = 0.51, Figure 2D) and a clean and visually discrete pattern of points consistent with CN signal from SNP6 data. These results suggest that the genomic background learned by `dryclean` can help reveal CN signal in high-density read depth data that is otherwise obscured by assay– and batch-specific bias.

WGS waves reflect DNA replication. Though WGS theoretically comprises uniformly sampled fragments from a DNA extract, cancer WGS data often contain read depth oscillations of low frequency ($\geq$ 100kbp) and varying amplitude, also termed "waves". These waves, also observed in microarrays (19, 20), cause false positive calls in downstream segmentation algorithms that detect SCNA as read depth change points. WGS waviness is most commonly attributed to technical factors (e.g. GC bias) during library preparation (10). Previous work in non-malignant cells have also linked fluctuations in WGS read depth to the ongoing replication of cells (21). A key challenge in detecting replication driven fluctuations is separating them from SCNAs, which are abundant in aneuploid cancer cells.

To examine the relationship between `dryclean` output, copy number, and replication timing we analyzed a particularly "wavy" WGS-profiled CCLE cell line (G16478.KE-39, stomach cancer cell line (22)). We compared read depth signal before and after `dryclean` correction to an average of 6 lymphoblastoid cell line (LCL) replication timing profiles obtained from two 1KGP trios (21). Indeed, analysis of 1000 randomly selected genomic bins of the `dryclean` input showed non-negligible correlation (R^2 = 0.16, Figure 3A) between replication timing z-scores and GC– / mappability-corrected read depth. This correlation was complicated by the presence of SCNAs, which defined discrete clusters of genomic bins (Figure 3A, colors represent CN from SNP6-derived segments). This replication timing correlation disappeared when comparing against the `dryclean`-processed read depth foreground ($\mathbf{f}$) vector (R^2 = 0.002, Figure 3B), where the colored data points separated primarily on the basis of copy number. In contrast, the `dryclean`-derived background ($\mathbf{b}$) vector showed high correlation with replication timing (R^2 = 0.45, Figure 3C) and absence of correlation with SNP6-derived copy number. The ability of `dryclean` to separate the GC-corrected input signal into an SCNA-enriched foreground and replication timing-enriched background was further demonstrated in Figure 3D, where the background vector tracked with LCL replication timing while the foreground change points co-located with concordantly oriented rearrangement junctions.

Expanding this analysis to a larger dataset, we observed variable correlation (R^2 range from 0.2 to 0.6) between background $\mathbf{b}$ vectors computed for 320 CCLE cell lines (22) and the LCL replication profile reference (Supplementary Figure S1). To quantify the magnitude of replication-associated waviness, we computed a "replication score" as the covariance of $\mathbf{b}$ and LCL replication timing vectors and classified cell lines as high, moderate or low replicating. Figure 3E demonstrates variable degrees of `dryclean`-derived background $\mathbf{b}$ vector waviness and its correlation with LCL replication timing in an example region of chromosome 2p for high, moderate and low replicating CCLE lines.

To determine whether cell-line specific differences in replication-correlated waviness represent differences in S-phase fraction in cancer cell lines, we assessed the expression of the "meta-PCNA gene module" (23). This module comprises the top 1% of genes that are most positively correlated with the proliferating cell nuclear antigen (PCNA) expression across an independent set of 36 normal tissues (23). PCNA is an essential component of replication machinery and is a commonly used marker of proliferation. Figure 3F shows a significant increase in PCNA module expression between low, moderate, and high-replicating CCLE groups (Wilcoxon rank sum test p value $= 1.9 \times 10^{-7}$ between high and low, p value $= 6.47 \times 10^{-5}$ between moderate and low and p = 0.01933 between high and moderate).

`dryclean` reduces bias and improves SCNA detection. We broadly evaluated `dryclean`'s ability to reduce bias in read depth data across a dataset of WGS, WES and simulated TP. Comparing two metrics of noise and waviness (DLRS, hypersegmentation, see Methods for details) across 1576 samples and two platforms (WES and TP), we saw significant and consistent reduction in both metrics for `dryclean`-processed read depth relative to TNR (Wilcoxon rank-sum test p value $< 2.2 \times 10^{-16}$ for all comparisons) (Supplementary Figure S2B-C). Moreover, for both WES and TP data, we identified a subset of TNR outliers that were "rescued" by `dryclean` (Supplementary Figure S2A). We

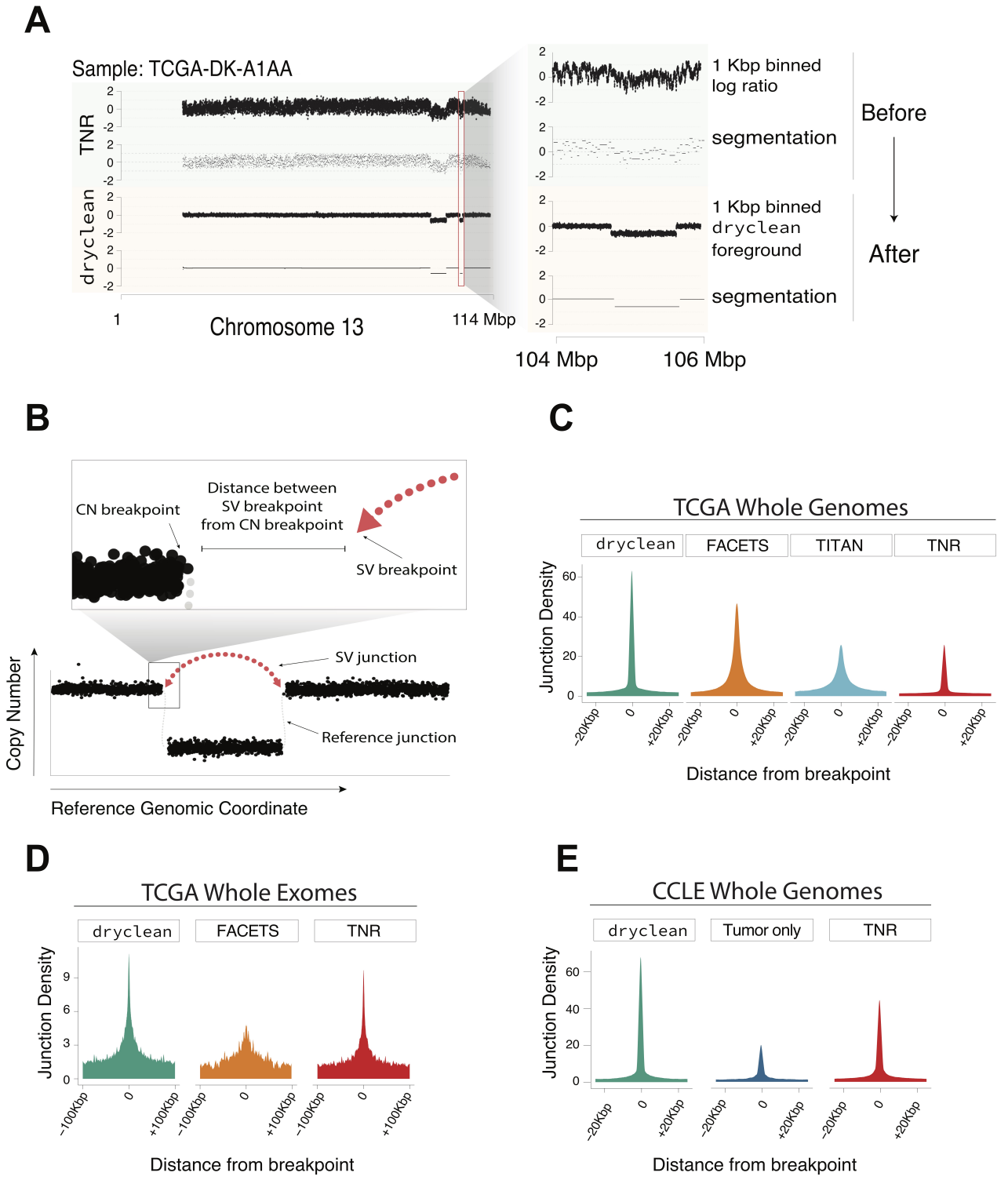


Fig. 4. `dryclean` improves SCNA signal in WGS and WES read depth data. (A) Example of a TCGA WGS case with noisy / wavy TNR and hypersegmented CBS output. The `dryclean` inferred foreground demonstrates uniform coverage except for CN endpoints, yielding a dramatically improved segmentation. (B) Schematic of analyses in C-E, which rely on the premise that structural variant junction break points should occur at nearly all true SCNA. (C-E) Distribution of structural variation junctions around CN endpoints from TNR, FACETS, `dryclean` and TITAN in (C) TCGA WGS, (D) TCGA WES and (E) CCLE WGS data. CN end-points from `dryclean` segmentations have the highest and most focal density of junctions within 20 Kbp (WGS) and within 100 Kbp (WES) relative to other methods.

observed similar reduction of noise metrics relative to TNR across 943 WGS cases (Supplementary Figure S2B-C). Figure 4A shows chromosome 13 of a bladder cancer (TCGA-DK-A1AA) WGS sample with very noisy coverage (DLRS= 0.145) and hypersegmentation that persisted despite GC bias and mappability correction. Following `dryclean` coverage correction, the DLRS value was lowered to 0.06 and SCNAs emerged visibly in the read depth data and subsequent seg-

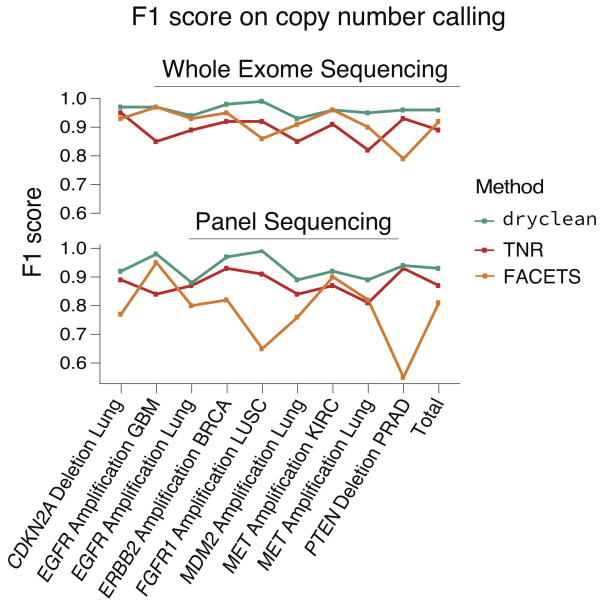


Fig. 5. `dryclean` enables superior detection of clinically actionable SCNA. Upper panel shows F1 scores of different methods with respect to SNP6. `dryclean` shows the highest accuracy, with F1 scores exceeding 90% for all loci and cancer types. The lower panel shows the F1 score for TP. Lung includes lung adenocarcinoma and lung squamous cell cancers.

mentation.

Since `dryclean` is not an SCNA caller *per se* but a method for reducing read depth bias, we combined it with a known segmentation method called circular binary segmentation (CBS, (24)). We compared the resulting SCNA calls (CBS-DC) against FACETS (14) or CBS combined with TNR (CBS-TNR), two representative state-of-the-art approaches. Using SNP6 segmentation as an orthogonal gold standard, we found that the locations of CBS-DC segment endpoints were significantly more concordant with SNP6 than CBS-TNR and FACETS results in both WES and TP data (Wilcoxon rank-sum test p value < 0.001 for all comparisons, Supplementary Figure S3A). In addition, the CBS-DC copy status (loss vs. neutral vs. gain, see Methods for details) at overlapping segments showed greater concordance with SNP6 than CBS-TNR and FACETS (Wilcoxon rank sum test p value $< 2.2 \times 10^{-16}$ for all comparisons) (Supplementary Figure S3B).

Since WGS can theoretically provide a more accurate estimate of CN than microarrays, SNP6 data does not provide an adequate "gold standard" for WGS-derived calls. Thus, to evaluate `dryclean`'s WGS performance, we used the presence of structural variant (SV) junctions (Figure 4B) as an orthogonal benchmark of SCNA accuracy.

Specifically, we quantified the distance between SvABA-derived (25) rearrangement junctions and SCNA segment end-points generated by `dryclean` vs. three state-of-the-art SCNA methods (CBS-TNR, FACETS, TITAN) (13, 14). Indeed, we found that `dryclean`-derived SCNA were focally enriched in SvABA junctions within a few Kbp of the segment end point (Figure 4C). We saw similar improvement in WES, where SV junctions were most highly enriched around `dryclean` breakpoints (Figure 4D) relative to other state-of-the-art approaches, including FACETS. Finally, application of `dryclean` to 320 CCLE cancer cell lines (22) lacking a matched normal demonstrated improved SCNA detection relative to two versions of CBS (FACETS and TITAN could not be compared, since they require a matched normal) (Figure 4E, see Methods). These results suggest that the read depth noise reduction achieved by `dryclean` enriches for signal associated with true SCNA across multiple sequencing platforms and sample types.

`dryclean` provides superior detection of clinically actionable SCNA. We assessed `dryclean`'s ability to detect clinically-actionable SCNAs, i.e. those potentially informing prognosis and/or treatment either as standard of care or in the context of a clinical trial. We defined actionable events using annotations from Memorial Sloan Kettering Cancer Center's OncoKB database (26) with the addition of some genes known to be prognostic in a particular cancer type (*CDKN2A* deletion (27) and *MDM2* amplification (28) in non-small cell lung cancer, *PTEN* deletion (29) in prostate adenocarcinoma). We computed precision and recall using the SNP6 annotation as the gold standard for both WES and TP analyses. As shown in Figure 5, F1 scores (harmonic mean of precision and recall) for `dryclean` matched or exceeded those for FACETS and TNR for several clinically relevant CNA for both WES and TP. Table S1 shows the sample counts used for the F1 calculations.

The presence of WES and TP bias obscured clinically important SCNA and resulted in important false negatives (Supplementary Figure S4A-C). This included focal *PTEN* loss and *MET* amplification, each of which was missed by TNR and FACETS. Supplementary Figure S4C provides a cautionary vignette on the use of matched normal samples in SCNA calling. Instead of calling a focal amplification in *EGFR*, which were seen in both `dryclean` and SNP6, both TNR and FACETS identified an artifactual deletion. Inspection of raw read counts in the matched normal sample (Supplementary Figure S5) showed that the "solid tissue normal" sample also had a broad and high amplitude *EGFR* amplification, possibly indicating a sample mix up or tumor-in-normal contamination. When this sample was used as a paired normal, it created a spurious drop in signal that was interpreted by downstream tools as a deletion. By modeling a broad PON, `dryclean` was immune to this artifact and agreed exactly with the SNP6 signal.

`dryclean` detects SCNA in dilute tumor samples. The detection of somatic variants in circulating tumor DNA (ctDNA) enables non-invasive monitoring of relapse and resistance in patients receiving targeted drugs and/or chemotherapy (5, 30, 31). `dryclean`'s ability to enrich SCNA signal above an assay– and batch-specific background could be potentially leveraged to detect microscopic relapse or drug resistant driver genotypes in WGS of ctDNA or solid biopsies. SCNA-based relapse detection might be particularly sensitive in cases where a primary tumor SCNA profile was previously obtained with WGS, WES, or TP.

To explore this application for breast cancer relapse de-

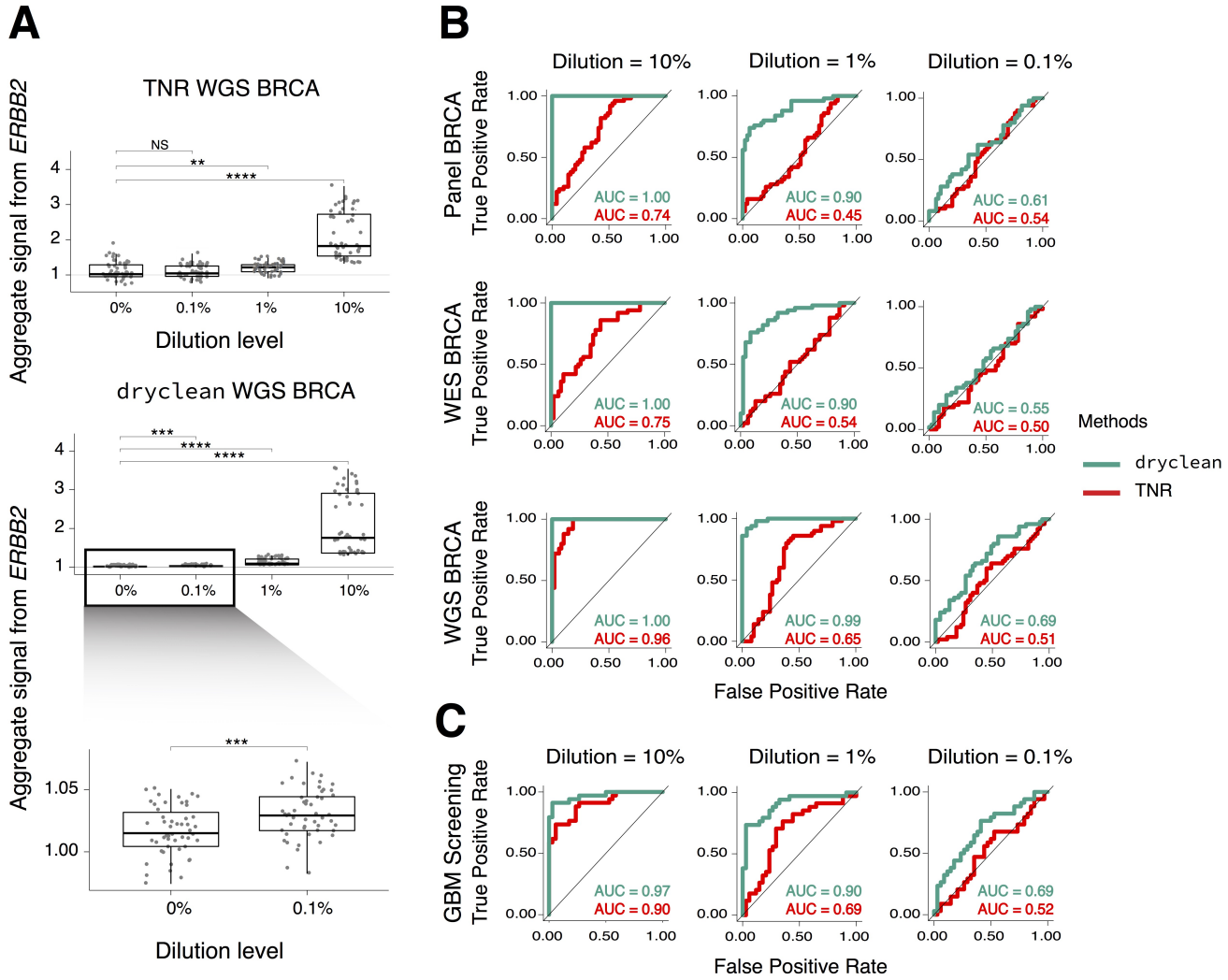


Fig. 6. `dryclean` enables sensitive detection of dilute tumor DNA. (A) Aggregated intensity of `dryclean` foreground vs. TNR signal at the *ERBB2* locus is shown across various *in silico* dilutions of tumor in blood WGS. `dryclean` detects *ERBB2* amplification at as low as 0.1% tumor concentration (lower panel, zoomed), relative to control dilutions with 0 tumor DNA or results obtained from patients WT for *ERBB2*. TNR-based SCNA signal is indistinguishable from negative controls at 0.1% dilution. (B) Receiver operating characteristic (ROC) curves for `dryclean` and TNR demonstrate increased sensitivity *in silico* dilutions of tumor DNA from *ERBB2* amplified breast cancers, with highest improvement and sensitivity in WGS. (C) *in silico* dilution results from 34 randomly selected GBM WGS samples for which SCNA signal was aggregated across 34 loci known to be recurrently amplified in GBM. These results show a similar improvement in tumor DNA detection for `dryclean` relative to other approaches for as low as 0.1% dilution.

tection, we generated 30X tumor-normal *in silico* dilutions from five TCGA breast cancer cases with amplified *ERBB2* for which we had WGS, WES, and simulated TP data. We examined 5 dilution fractions (10%, 1%, 0.1%, 0.01% and 0%) in 10 replicates for changes in mean read depth at the *ERBB2* locus relative to pure germline DNA (0% mixture where the paired normal sample was diluted into a randomly selected blood sample).

Figure 6A depicts the signal from TNR and `dryclean`. We were able to detect significant signal in `dryclean` output at *ERBB2* for as low as 1 in 1000 tumor cell concentration (Wilcoxon rank sum test p value $= 5.5 \times 10^{-4}$). In contrast, TNR was unable to detect significant signal relative to negative control samples at lower than 1 in 100 tumor DNA fraction, below which changes in signal intensity were indistinguishable from noise. Receiver operating characteristic (ROC) curves (Figure 6B) demonstrated superior tumor detection with `dryclean` relative to TNR across various dilutions and sequencing assays.

Specifically, `dryclean` outperformed TNR across platforms for dilutions as low as 1 in 1000 in WGS. This suggests that `dryclean` (in combination with a primary tumor profile and a suitable PON) can significantly improve the detection of microscopic burdens of tumor cells and clinically relevant SCNAs in relapse tissue.

To further explore the utility of `dryclean` in detecting ctDNA in the absence of a primary tumor profile, we carried out *in silico* dilution experiments on 34 GBM WGS samples from TCGA. For tumor detection, we chose a panel of previously described (32) frequently amplified genes in GBM as markers across which we aggregated `dryclean` and TNR signal. We found that `dryclean` was able to reliably detect the presence of tumor at as low as 1% dilution (area under curve, AUC = 0.90), greatly exceeding the performance of

TNR (AUC = 0.69) (Figure 6C). These results suggests that dryclean enrichment of foreground SCNA signal can improve tumor detection even in the absence of a primary tumor WGS profile.

Discussion

We have demonstrated that dryclean improves the accuracy of SCNA detection across multiple sequencing platforms. However, we note that dryclean is not a method for SCNA characterization, but an algorithm for read depth normalization. In this study, we combine dryclean with CBS (CBS-DC) (24), primarily due to ease of implementation. Additional performance gains over CBS-DC might be obtained by combining dryclean with other, more sophisticated, SCNA detection algorithms (e.g. TITAN, FACETS) (13, 14) through minor bioinformatic modifications that replace the TNR input of these methods with the dryclean foreground signal.

The results obtained from dryclean greatly depend on the composition of the PON, which ideally should reflect the spectrum of assay and batch specific bias that is observed in a new tumor sample. Increasing the number of normal samples in the PON will incrementally improve the quality of results (Supplementary Figure S6, see Methods), but larger PONs require more compute to train (in the batch learning step (Figure 1)). Given standard computational limitations (storage, RAM, CPU), it is useful in practice to build a PON of 50-500 samples that is balanced with respect to common noise modes. If the PON is not representative of the assay and batch patterns in the tumor samples being analyzed, dryclean will produce sub-optimal results. This is especially true if the PON and tumor samples are derived from different sequencing platforms (e.g. HiSeq 2500 vs. HiSeq 4000 vs. NovaSeq). A notable use case of dryclean is in the analysis of cancer cell lines for which a matched normal can't be found but for which a reasonable PON exists. We did not explore applying dryclean to the analysis of germline CN in this study; however, we believe that it may prove similarly useful in the sensitive detection of rare constitutional CN variants in clinical samples and populations.

Our dryclean analyses of cancer WGS implicate replication timing in the commonly noted "wave artifacts" in cancer sequencing data. Our analyses raise the possibility of using dryclean to chart the landscape of replication timing in cancer samples, including the nomination of somatic replication timing quantitative trait loci (21). However, since the dryclean background is inferred from a panel of normals, a replication timing-inference pipeline that depends exclusively on this signal may miss important replication timing alterations that arise in lineages that are not represented in the PON or result from large-scale somatic rearrangements. We foresee that additional processing downstream of dryclean foreground and background signals will be required to derive an optimally accurate replication timing signal from cancer WGS.

Our WGS dilution results suggest that dryclean may help dramatically increase the sensitivity of SCNA detection in the setting of cancer screening or relapse monitoring, including solid biopsy as well as liquid biopsy of cell free plasma or peripheral blood (e.g. clonal hematopoiesis) (5, 7, 31). Since previously published ctDNA methods can only reliably detect tumor-specific variants at an allelic or tumor-cell fraction > 0.1 (WES, WGS), our results suggest that dryclean may increase the sensitivity of ctDNA tumor detection as much as 10-fold over state-of-the-art approaches (6, 31). Since our *in silico* dilutions were performed on solid tissue-derived sequencing data, additional studies will be required to validate the utility of dryclean on liquid biopsy. This will include the generation and application of a ctDNA PON that captures the unique library characteristics of this assay (e.g. fragmentation patterns (33)).

Though we have demonstrated the use of dryclean on WES, TP, and WGS data from TCGA, we believe dryclean can improve the detection of SCNA across a variety of sequencing platforms, not just standard DNA sequencing. Future applications of dryclean may also include single-cell assays (e.g. scCNV, scATAC) that produce thousands of genomic profiles in a single experiment but may suffer from significant read depth bias (e.g. whole genome amplification, transposase accessibility) as well as sparsity (34). Application of dryclean in these data may help to better characterize genomic heterogeneity and link single cell SCNA patterns to phenotypic readouts of cell state.

Conclusion

Read depth bias limits the accuracy of SCNA detection in clinical and research cancer DNA sequencing. In this study, we have introduced and benchmarked dryclean, a method to remove assay– and batch-specific bias from read depth data across a several sequencing platforms (WES, TP, WGS). These results indicate that dryclean could be a powerful addition to TP, WES, and/or WGS clinical oncology bioinformatics pipelines. Analysis of dryclean-derived background indicate that biological (i.e. replication timing), rather than technical factors (i.e. GC bias during library preparation) drive the "wave artifact" commonly observed in cancer WGS. Analysis of *in silico* tumor dilutions simulating solid or liquid biopsy sequencing data indicates a potential role for dryclean in sensitive relapse detection.

Methods

Mathematical formulation of dryclean . A basic premise of dryclean is that assay and batch-specific bias define a low-rank subspace in high-dimensional read depth data, which we represent as a matrix of m genomic bins across n samples. We achieve this through the application of robust principal component analysis (rPCA), which can accurately separate low-rank structures and high-amplitude outliers in complex multivariate data (18). A natural application for rPCA is motion detection in video data, where each data vector represents a video frame of a figure moving against a mostly stable background.

rPCA decomposes an input matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ into a

low-rank matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$ and sparse matrix $\mathbf{F} \in \mathbb{R}^{m \times n}$ through the solution of a constrained optimization described in Eq. 1. This can be solved using iterative optimization methods such as Accelerated Proximal Gradient (APG) or Augmented Lagrangian Multiplier (ALM). Details of these algorithms can be found in (35, 36) respectively.

In general, these approaches minimize the objective function:

$$\underset{\mathbf{B},\mathbf{F}}{\text{minimize}} \frac{1}{2} \|\mathbf{M} - \mathbf{B} - \mathbf{F}\|_F^2 + \lambda_1 \|\mathbf{B}\|_* + \lambda_2 \|\mathbf{F}\|_1 \quad \textbf{(2)}$$

where λ_1,λ_2 are tuning parameters set to $\lambda_1 = \lambda_2 = \frac{1}{\sqrt{max(m,n)}}$, and $||A||_F = (\sum_i^n \sum_j^m a_{ij}^2)^{\frac{1}{2}}$ is the Frobenius norm of matrix $\mathbf{A}$.

The robustness of rPCA comes at some computational cost. Training of Eq. 1 requires in-memory access to the full data matrix at every iteration of the optimization. This limits performance on larger data (35), e.g. read depth matrices comprising millions of bins and thousands of samples. This "batch" inference is also inefficient for the analysis of new data (e.g. a new tumor sample), which requires the model to be repeatedly refit to previously analyzed data points. To address this issue, we employ the online rPCA (OR-PCA) approach introduced by Feng et al (37), which allows efficient inference of additional data vectors, given an initially trained matrix.

`dryclean` pipeline. We combine the batch and online rPCA approaches in the `dryclean` pipeline (Figure 1). The first step in the pipeline is to infer a low-rank subspace from a PON read depth data. Let $\mathbf{D} \in \mathbb{R}^{m \times n}$ denote a matrix of log transformed read depth data across n samples and m bins, where each bin represents an exon, amplicon, hybrid capture probe, or tiled partition of the genome.

Applying the rPCA decomposition in Eq. 1 to $\mathbf{D}$, we infer a low-rank matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$ and a sparse matrix $\mathbf{F} \in \mathbb{R}^{m \times n}$. $\mathbf{B}$ is a low-rank representation of the PON, which comprises assay– and batch-specific noise (e.g. hybridization probe bias) but also can include common germline copy number polymorphisms (CNPs). The $\mathbf{F}$ matrix represents germline CN variants that are private to a few samples.

In practice, for inference, we use an online implementation that approximates the solution to (Eq. 1) for each tumor by projecting the tumor onto the subspace spanned by the n columns of the $\mathbf{B}$ and $\mathbf{F}$ matrix computed for just the PON. The latter then only needs to be computed once during a more (time and computationally intensive) training step.

`dryclean` is executed in three steps: First, read counts from a sample `.bam` are binned and corrected for GC content and mappability biases using a 2-step loess regression, in an implementation similar to HMMcopy (17). In the second step, a user defined number of normal samples are used to create a PON and extract the $\mathbf{B}$ matrix. Finally, using the $\mathbf{B}$ matrix, SCNA signals (foreground) for each tumor are extracted and stored as a Genomic Ranges object. Optionally, `dryclean` will use all normal samples to identify markers that commonly exhibit germline variation in a given dataset (Figure 1).

Though `dryclean` is a method for normalization of high-density read depth data, we also provide segmented output through the application of circular binary segmentation (CBS) to the $\mathbf{f}$ vector. We use these `dryclean`-derived SCNA calls throughout the manuscript to compare the quality of `dryclean` results to state-of-the-art methods; however, we foresee that `dryclean` output could be fed to more sophisticated segmentation algorithms, e.g. those employing additional data modalities such as heterozygous SNP counts or structural variant junctions to map SCNA with even higher accuracy and sensitivity. The contents of $\mathbf{f}$ vector represent the "foreground" of read depth data and captures the SCNAs while the $\mathbf{b}$ vector captures biases such as batch effect assay biases and replication timing which we term as "background". As is shown in subsequent sections both vectors provide insight into different aspects of tumor biology.

PON optimization. `dryclean`'s performance depends on the choice and number of normal (i.e. euploid) samples used to create the $\mathbf{B}$ matrix (PON). To objectively assess this performance, we use a measure of read depth noise called derivative log ratio spread (DLRS). It is a standard measure of the quality of log ratio (logR) data in SCNA analysis. It gauges marker to marker consistency in the signal and a larger spread is indicative of poor quality logR data (38). Applying DLRS across our benchmark analyses, we see that noise decreases as the PON size increases. Notably, we see a significant reduction in noise compared to TNR with a PON composed of as few as 50 samples (Supplementary Figure S6).

`dryclean` provides two options for sample selection for users wishing to generate smaller PONs from a larger dataset of normal samples. The first option is to build a PON from a randomly selected set of normal samples. A drawback of this approach is that less frequent batch effect modes may not be represented in the PON and thus `dryclean` will fail to correct coverage in tumor samples affected by those modes. Though one can build a very large and representative PON, the computational cost in building and applying such a dataset (even in batch rPCA mode) may be prohibitive. To address this issue, we provide users with a second option to create a PON that is balanced across batch effect modes. This is achieved efficiently by clustering on the basis of the low-rank signal ($\mathbf{b}$ vectors) computed for an rPCA decomposition performed across a small subset of genomic regions and a full set of normal samples and clustering these vectors. The normal samples to be used in the PON are then selected so that each cluster is represented. We see that there is rich structure among the normal samples (Supplementary Figure S7). This allows users to keep the size of the PON as small as possible while maximizing the information in the PON.

For all WES and TP `dryclean` analyses reported in the manuscript, we used a PON of 100 samples selected by the hierarchical clustering approach. To assess the performance of `dryclean` on WGS data (CCLE and TCGA), we generated a PON of 400 normal samples from the TCGA WGS cohort using the clustered approach

CNV filter. Since `dryclean` uses a PON and not individual paired normal samples, an additional step is required to separate the SCNA signal from common germline CN polymorphisms (CNP's) or matched normal germline CN variants (CNV's). `dryclean` performs this step after creating the PON. Each normal sample is decomposed against the PON and the $\mathbf{f}$ vectors are extracted. This vector captures the germline CNV / CNP as opposed the $\mathbf{b}$ vector that encodes the background. Supplementary Figure S8 shows a small region of the genome with all these vectors concatenated in column space. `dryclean` creates an $\mathbf{f}$ vectors for all normal samples and then, using user defined frequency thresholds, annotates the markers harboring germline CNV and CNP, which are removed from downstream SCNA recurrence (i.e. GISTIC 2.0 (39)) analysis. Supplementary Figure S9 shows the reduction in narrow peaks from GISTIC 2.0 after this filter is applied.

Data sources. Tumor and normal Illumina Hi-Seq `.bam` files were obtained from The Cancer Genome Atlas (TCGA, dbGAP identifier: phs000178.v1.p1) (40) via the GDC portal (https://portal.gdc.cancer.gov/). For WGS analyses, we used 943 samples with 23 different cancer types. For WES analysis, we used 1576 randomly selected samples comprising 19 different cancer types. We simulated targeted panel (TP) sequencing in tumor and normal samples by sub-sampling "target regions" from TCGA WES coverage data. We used 572 Cancer Gene Consensus (41) genes as targets. The distribution of the samples for different platforms can be found in (Supplementary Figure S10). 320 cancer cell line encyclopedia (CCLE) WGS were obtained from the CCLE portal (www.broadinstitute.org/ccle).

Benchmarking. We compared circular binary segmentation (CBS) (24) ran on `dryclean` (CBS-DC) against two commonly used approaches for SCNA calling, CBS ran on TNR (CBS-TNR), and FACETS (14). For CBS-TNR, we used log-transformed, GC and mappability corrected TNR as an input. For CBS-TNR in CCLE cell lines that lack a matched normal, we computed a "pseudo-TNR" whose denominator was a "composite diploid sample" obtained by averaging GC and mappability-corrected read depth profiles across 943 TCGA WGS normal samples.

FACETS employs allele-specific copy number analysis of reference and variant allele read counts for germline polymorphic markers annotated by dbSNP (200K markers for WES and 7,400 markers for simulated TP). To assess the quality of WES and TP analyses, we used results from Affymetrix Genome-Wide Human SNP Array 6.0 (SNP6). SNP6 and other high-resolution microarray technologies are routinely used in basic research and clinical practice to efficiently detect SCNAs and represent the gold standard in clinical cytogenetics labs (26).

For all copy number comparisons, we removed centromeric segments, which are not reliably called by any SCNA method, including SNP6. We also limited our analysis to segments of 1MB or larger in size, since the reliability of more focal SCNA calls on any platform (including SNP6) is imperfect (14). For segment concordance, we defined concordant segments as those with at least 70% reciprocal overlap between both the start and end positions in the subject and query. To assess CN concordance, we used standard thresholds on segment intensities (i.e. segment mean value for CBS-DC and CBS-TNR or "cnlr.median" field in FACETS) to assign segments a "CN status", i.e. gained vs lost. Specifically, we classified segments with intensities $>$ 0.1 as "gains" and < -0.1 as "losses". We then measured the proportion of correct "gain" vs. "loss" designations across all genomic coordinate-concordant segments (from above).

ctDNA WGS simulations. For breast cancer (BRCA) samples, we simulated WGS dilutions at 5 dilution levels ($\alpha =$ 10%, 1%, 0.1%, 0.01% and 0%), 10 replicates, and 5 cases for a total of 250 samples. For each sample, dilution α, and 1 Kbp bin i with tumor and peripheral blood read count t_i and p_i, we sampled spiked tumor counts $\hat{t}_i \sim Pois(\lambda_i)$ from a Poisson distribution with $\lambda_i = \alpha t_i \frac{||\mathbf{p}||_1}{||\mathbf{t}||_1}$. The simulated dilution read counts s_i for each bin i were then defined as $s_i = p_i + \hat{t}_i$. We then processed read counts through `dryclean` or TNR and analyzed read density at the *ERBB2* locus across various dilutions. For each simulation, we used a randomly chosen blood sample. Similar to BRCA, for glioblastoma (GBM) we used 34 samples and simulated dilutions at 5 dilution levels ($\alpha =$ 10%, 1%, 0.1%, 0.01% and 0%).

Replication timing analysis. We ran `dryclean` on 320 CCLE WGS samples. For each sample, the inferred background ($\mathbf{b}$ vector) was then collapsed at 1 Mb level and median smoothed. We calculated the replication "score" by taking the inner product of $\mathbf{b}$ vector with the lymphoblastoid cell line (LCL) replication timing profile (21), an average of 6 replication timing profiles derived from 1000 genomes LCLs. We then classified cell lines as high (score greater than 75th percentile), moderate (score between 75th and 25th percentile) and low replicating (score less than 25th percentile). Finally, we validated this classification by comparing the expression of genes associated with cell proliferation (23) within these groups.

Availability of data and software. `dryclean` is available in the GitHub repository https://github.com/mskilab/dryclean Data used to generate these results are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. (TCGA, db-GAP identifier: phs000178.v1.p1). All the CCLE processed datasets are available at the CCLE portal (www.broadinstitute.org/ccle) and DepMap portal (http://www.depmap.org). Raw sequencing data are available at Sequence Read Archive (SRA) under accession number PRJNA523380.

References

1. Dennis J Slamon, Brian Leyland-Jones, Steven Shak, Hank Fuchs, Virginia Paton, Alex Bajamonde, Thomas Fleming, Wolfgang Eiermann, Janet Wolter, Mark Pegram, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med*, 344(11):783–792, 2001.

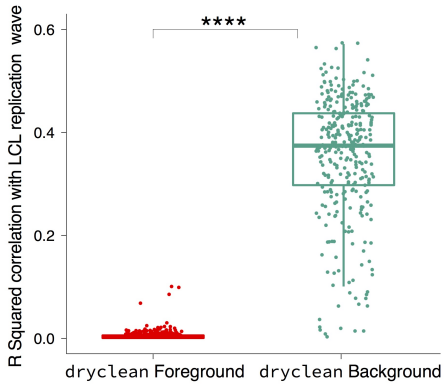
2. W Ruprecht Wiedemeyer, Ian F Dunn, Steven N Quayle, Jianhua Zhang, Milan G Chheda, Gavin P Dunn, Li Zhuang, Joseph Rosenbluh, Shujuan Chen, Yonghong Xiao, et al. Pattern of retinoblastoma pathway inactivation dictates response to CDK4/6 inhibition in GBM. *Proc. Natl. Acad. Sci. U. S. A.*, 107(25):11501–11506, 2010.
3. Matthew J Niederst, Lecia V Sequist, John T Poirier, Craig H Mermel, Elizabeth L Lockerman, Angel R Garcia, Ryohei Katayama, Carlotta Costa, Kenneth N Ross, Teresa Moran, et al. RB loss in resistant EGFR mutant lung adenocarcinomas that transform to small-cell lung cancer. *Nat. Commun.*, 6:6377, 2015.
4. David M Hyman, Barry S Taylor, and José Baselga. Implementing Genome-Driven oncology. *Cell*, 168(4):584–599, February 2017.
5. Luis A Diaz, Jr and Alberto Bardelli. Liquid biopsies: genotyping circulating tumor DNA. *J. Clin. Oncol.*, 32(6):579–586, February 2014.
6. Viktor A. Adalsteinsson, Gavin Ha, Samuel S. Freeman, Atish D. Choudhury, Daniel G. Stover, Heather A. Parsons, Gregory Gydush, Sarah C. Reed, Denisse Rotem, Justin Rhoades, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications*, 8:1324, 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-00965-y.
7. Po-Ru Loh, Giulio Genovese, Robert E Handsaker, Hilary K Finucane, Yakir A Reshef, Pier Francesco Palamara, Brenda M Birmann, Michael E Talkowski, Samuel F Bakhoum, Steven A McCarroll, and Alkes L Price. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature*, 44:651 355, 2018. ISSN 0028-0836. doi: 10.1038/s41586-018-0321-x.
8. Jin-Xing Liu, Yu-Tian Wang, Chun-Hou Zheng, Wen Sha, Jian-Xun Mi, and Yong Xu. Robust PCA based method for discovering differentially expressed genes. *BMC Bioinformatics*, 14 Suppl 8:S3, 2013.
9. Eric Samorodnitsky, Jharna Datta, Benjamin M Jewell, Raffi Hagopian, Jharna Miya, Michele R Wing, Senthilkumar Damodaran, Juliana M Lippus, Julie W Reeser, Darshna Bhatt, et al. Comparison of custom capture for targeted next-generation DNA sequencing. *J. Mol. Diagn.*, 17(1):64–75, 2015.
10. Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40:e72–e72, 2012. ISSN 0305-1048. doi: 10.1093/nar/gks001.
11. Angel Janevski, Vinay Varadan, Sitharthan Kamalakaran, Nilanjana Banerjee, and Nevenka Dimitrova. Effective normalization for copy number variation detection from whole genome sequencing. *BMC Genomics*, 13 Suppl 6:S16, 2012.
12. Ruibin Xi, Semin Lee, Yuchao Xia, Tae-Min Kim, and Peter J Park. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic acids research*, 44:6274 6286, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw491.
13. Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, 24(11):1881–1893, 2014.
14. Ronglai Shen and Venkatraman E Seshan. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.*, 44 (16):e131, 2016.
15. Fatima Zare, Michelle Dow, Nicholas Monteleone, Abdelrahman Hosny, and Sheida Nabavi. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, 18:286, 2017. ISSN 1471-2105. doi: 10.1186/s12859-017-1705-x.
16. Morgane Pierre-Jean, Guillem Rigaill, and Pierre Neuvial. Performance evaluation of DNA copy number segmentation methods. *Briefings in Bioinformatics*, 16:600–615, 2015. ISSN 1467-5463. doi: 10.1093/bib/bbu026.
17. D Lai and S Shah. HMMcopy: copy number prediction with correction for GC and mappability bias for HTS data. *R package version*, 1(0), 2012.
18. Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011.
19. John C. Marioni, Natalie P. Thorne, Armand Valsesia, Tomas Fitzgerald, Richard Redon, Heike Fiegler, T. Daniel Andrews, Barbara E. Stranger, Andrew G. Lynch, Emmanouil T. Dermitzakis, Nigel P. Carter, Simon Tavaré, and Matthew E. Hurles. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biology*, 8(10):R228, Oct 2007. ISSN 1474-760X. doi: 10.1186/gb-2007-8-10-r228.
20. Liangcai Zhang and Li Zhang. Use of autocorrelation scanning in DNA copy number analysis. *Bioinformatics*, 29(21):2678–2682, 09 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt479.
21. Amnon Koren, Robert E Handsaker, Nolan Kamitaki, Rosa Karlić, Sulagna Ghosh, Paz Polak, Kevin Eggan, and Steven A McCarroll. Genetic variation in human DNA replication timing. *Cell*, 159(5):1015–1026, November 2014.
22. Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald, 3rd, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, Kevin Hu, Alexander Y Andreev-Drakhlin, Jaegil Kim, Julian M Hess, Brian J Haas, François Aguet, Barbara A Weir, Michael V Rothberg, Brenton R Paolella, Michael S Lawrence, Rehan Akbani, Yiling Lu, Hong L Tiv, Prafulla C Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehar, Joshua M Korn, Dale A Porter, Michael D Jones, Javad Golji, Giordano Caponigro, Jordan E Taylor, Caitlin M Dunning, Amanda L Creech, Allison C Warren, James M McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E Maruvka, Andrew D Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D Jaffe, Andrew A Lane, David M Weinstock, Cory M Johannessen, Michael P Morrissey, Frank Stegmeier, Robert Schlegel, William C Hahn, Gad Getz, Gordon B Mills, Jesse S Boehm, Todd R Golub, Levi A Garraway, and William R Sellers. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, May 2019.
23. David Venet, Jacques E Dumont, and Vincent Detours. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.*, 7 (10):e1002240, October 2011.
24. Adam B Olshen, E S Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4): 557–572, 2004.
25. Jeremiah A Wala, Pratiti Bandopadhayay, Noah F Greenwald, Ryan O'Rourke, Ted Sharpe, Chip Stewart, Steve Schumacher, Yilong Li, Joachim Weischenfeldt, Xiaotong Yao, et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.*, 28(4):581–591, 2018.
26. D Chakravarty, J Gao, S M Phillips, R Kundra, H Zhang, J Wang, J E Rudolph, R Yaeger, T Soumerai, M H Nissan, et al. OncoKB: a precision oncology knowledge base. *Precis Oncol*, 2017.
27. Weiqiang Zhao, Cheng C Huang, Gregory A Otterson, Marino E Leon, Yan Tang, Konstantin Shilo, and Miguel A Villalona. Altered p16(INK4) and RB1 expressions are associated with poor prognosis in patients with nonsmall cell lung cancer. *J. Oncol.*, 2012:957437, April 2012.
28. M Higashiyama, O Doi, K Kodama, H Yokouchi, T Kasugai, S Ishiguro, K Takami, T Nakayama, and I Nishisho. MDM2 gene amplification and expression in non-small-cell lung cancer: immunohistochemical expression of its protein is a favourable prognostic marker in patients without p53 protein accumulation. *Br. J. Cancer*, 75(9):1302–1308, 1997.
29. Tamara L Lotan, Bora Gurel, Siobhan Sutcliffe, David Esopi, Wennuan Liu, Jianfeng Xu, Jessica L Hicks, Ben H Park, Elizabeth Humphreys, Alan W Partin, Misop Han, George J Netto, William B Isaacs, and Angelo M De Marzo. PTEN protein loss by immunostaining: analytic validation and prognostic indicator for a high risk surgical cohort of prostate cancer patients. *Clin. Cancer Res.*, 17(20):6563–6573, October 2011.
30. Makoto Nakabayashi, Akihiro Kawashima, Rika Yasuhara, Yosuke Hayakawa, Shingo Miyamoto, Chiaki Iizuka, and Akihiko Sekizawa. Massively parallel sequencing of cell-free DNA in plasma for detecting gynaecological tumour-associated copy number alteration. *Sci. Rep.*, 8(1):11205, 2018.
31. Klaus Pantel and Catherine Alix-Panabières. Liquid biopsy and minimal residual disease — latest advances and implications for cure. *Nature Reviews Clinical Oncology*, page 1, 2019. ISSN 1759-4782. doi: 10.1038/s41571-019-0187-3.
32. Cameron W Brennan, Roel G W Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty, J Zachary Sanborn, Samuel H Berman, Rameen Beroukhim, Brady Bernard, Chang-Jiun Wu, Giannicola Genovese, Ilya Shmulevich, Jill Barnholtz-Sloan, Lihua Zou, Rahulsimham Vegesna, Sachet A Shukla, Giovanni Ciriello, W K Yung, Wei Zhang, Carrie Sougnez, Tom Mikkelsen, Kenneth Aldape, Darell D Bigner, Erwin G Van Meir, Michael Prados, Andrew Sloan, Keith L Black, Jennifer Eschbacher, Gaetano Finocchiaro, William Friedman, David W Andrews, Abhijit Guha, Mary Iacocca, Brian P O'Neill, Greg Foltz, Jerome Myers, Daniel J Weisenberger, Robert Penny, Raju Kucherlapati, Charles M Perou, D Neil Hayes, Richard Gibbs, Marco Marra, Gordon B Mills, Eric Lander, Paul Spellman, Richard Wilson, Chris Sander, John Weinstein, Matthew Meyerson, Stacey Gabriel, Peter W Laird, David Haussler, Gad Getz, Lynda Chin, and TCGA Research Network. The somatic genomic landscape of glioblastoma. *Cell*, 155 (2):462–477, October 2013.
33. Stephen Cristiano, Alessandro Leal, Jillian Phallen, Jacob Fiksel, Vilmos Adleff, Daniel C Bruhm, Sarah Østrup Jensen, Jamie E Medina, Carolyn Hruban, James R White, Doreen N Palsgrove, Noushin Niknafs, Valsamo Anagnostou, Patrick Forde, Jarushka Naidoo, Kristen Marrone, Julie Brahmer, Brian D Woodward, Hatim Husain, Karlijn L van Rooijen, Mai-Britt Worm Ørntoft, Anders Husted Madsen, Cornelis J H van de Velde, Marcel Verheij, Annemieke Cats, Cornelis J A Punt, Geraldine R Vink, Nicole C T van Grieken, Miriam Koopman, Remond J A Fijneman, Julia S Johansen, Hans Jørgen Nielsen, Gerrit A Meijer, Claus Lindbjerg Andersen, Robert B Scharpf, and Victor E Velculescu. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, 570(7761):385–389, June 2019.
34. Guo-Cheng Yuan, Long Cai, Michael Elowitz, Tariq Enver, Guoping Fan, Guoji Guo, Rafael Irizarry, Peter Kharchenko, Junhyong Kim, Stuart Orkin, John Quackenbush, Assieh Saadatpour, Timm Schroeder, Ramesh Shivdasani, and Itay Tirosh. Challenges and emerging directions in single-cell analysis. *Genome Biol.*, 18(1):84, May 2017.
35. Wei Xiao, Xiaolin Huang, Jorge Silva, Saba Emrani, and Arin Chaudhuri. Online Robust Principal Component Analysis with Change Point Detection. *arXiv e-prints*, art. arXiv:1702.05698, Feb 2017.
36. Zhouchen Lin, Minming Chen, and Yi Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *arXiv e-prints*, art. arXiv:1009.5055, Sep 2010.
37. Jiashi Feng, Huan Xu, and Shuicheng Yan. Online robust PCA via stochastic optimization. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Adv Neural Inf Process Syst*, pages 404–412. Curran Associates, Inc., 2013.
38. Dalila Pinto, Katayoon Darvishi, Xinghua Shi, Diana Rajan, Diane Rigler, Tom Fitzgerald, Anath C Lionel, Bhooma Thiruvahindrapuram, Jeffrey R Macdonald, Ryan Mills, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.*, 29(6):512–520, 2011.
39. Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, 12(4): R41, 2011.
40. John N Weinstein, The Cancer Genome Atlas Research Network, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project, 2013.
41. John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, 47(D1):D941–D947, 2019.

Robust foreground detection in somatic copy number data

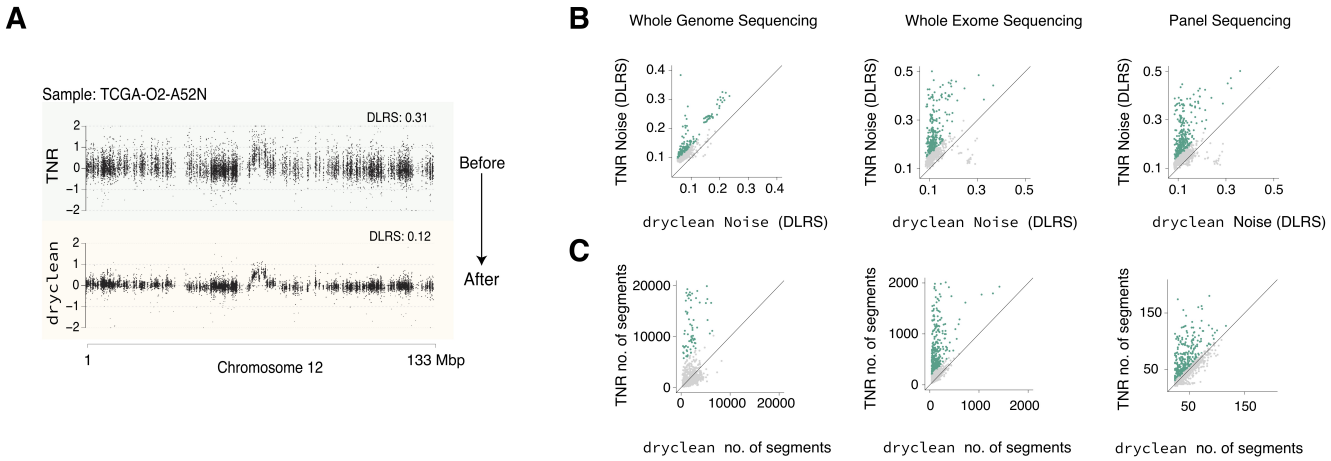
Supplementary File

Deshpande *et al.*

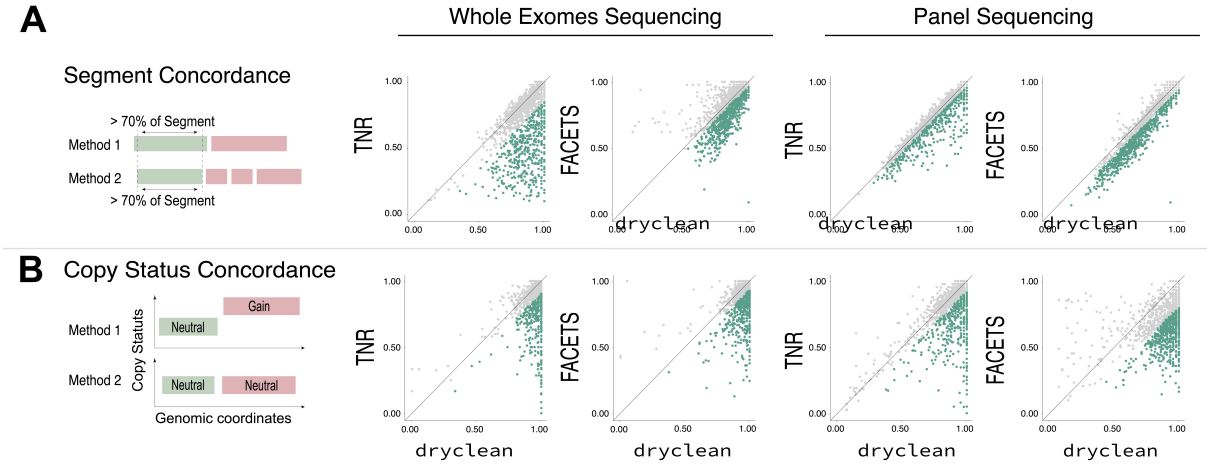
Supplementary Figures.



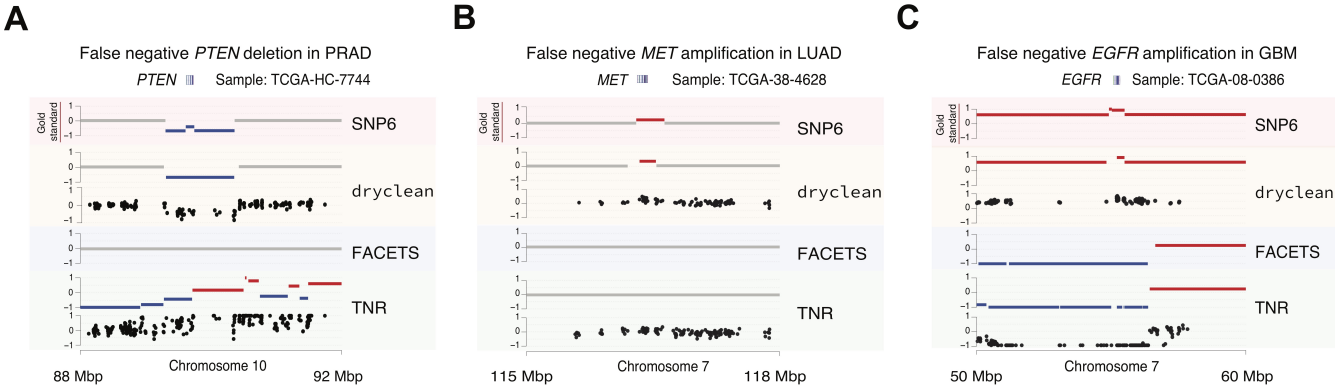
Supplementary Figure S1. Correlation of `dryclean` with LCL replication wave Higher correlation of `dryclean` background with LCL which is much higher than `dryclean` foreground that comprises of SCNAs. Each point here represents a CCLE cell line



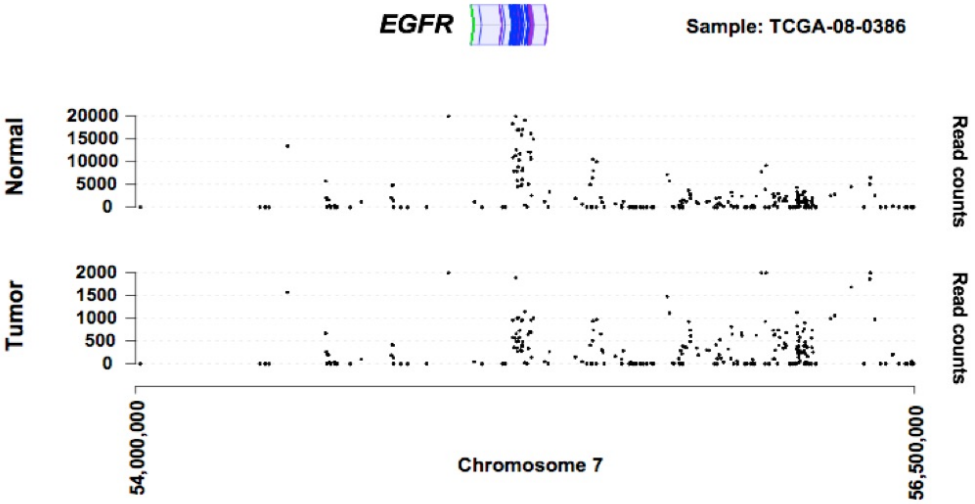
Supplementary Figure S2. `dryclean` reduces read depth noise in WES and TP. (A) A representative WES sample with decreased DLRS in `dryclean` (yellow panel) relative to TNR (green panel). The higher DLRS in the TNR indicates a high burden of assay and batch specific noise. (B) Pan-cancer comparison of DLRS between `dryclean` and TNR in WES and simulated TP across 1576 TCGA cases. Each data point represents a TCGA case. For both platforms, DLRS is significantly reduced by `dryclean` relative to TNR. (C) Pan-cancer comparison of hypersegmentation (number of CBS segments) shows a reduction in `dryclean` results relative to TNR for both WES and TP. Each data point represents a TCGA case.



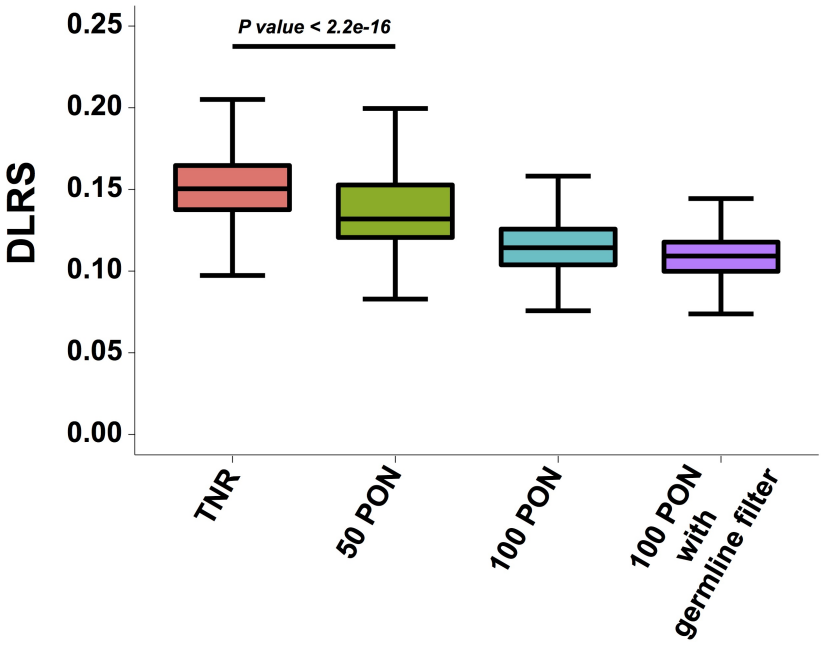
Supplementary Figure S3. Concordance with SNP6. (A) Proportion of concordant segments with respect to genomic position. dryclean shows higher number of concordant segments as compared to FACETS and TNR in WES and to a lower extent in TP. (B) proportion of concordant segments with respect to copy number signal is much higher in dryclean in both WES and TP.



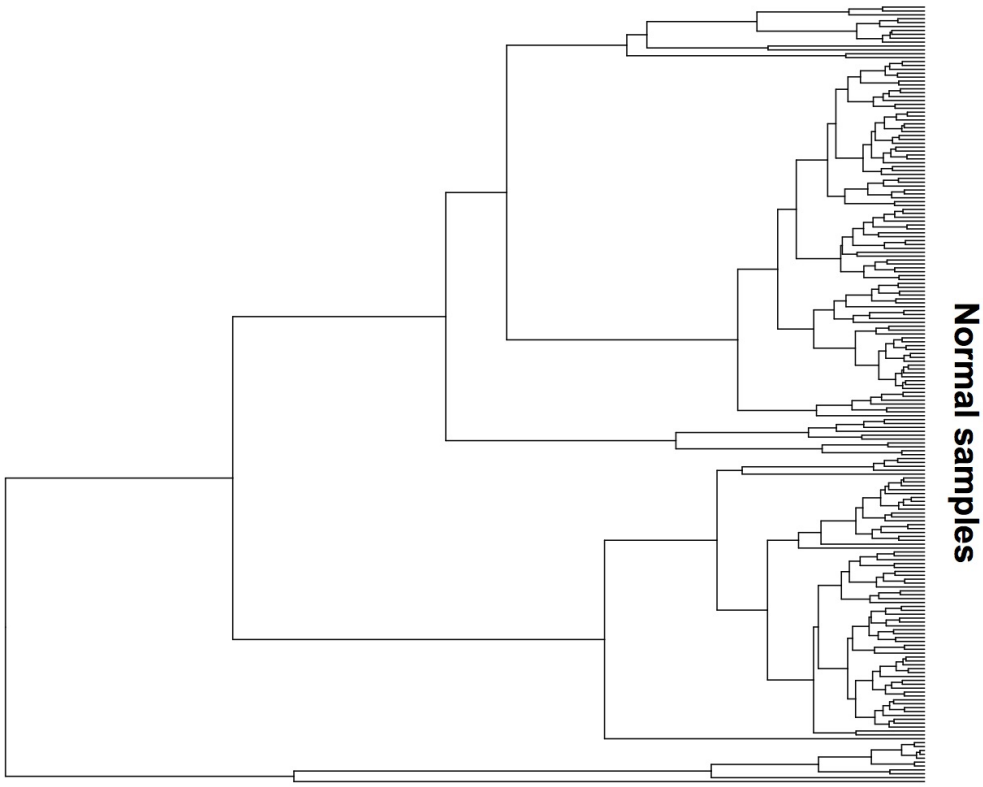
Supplementary Figure S4. False negative examples of clinically important SCNA.(A-C) show instances of `dryclean` rescue of FACETS and/or TNR false negative calls in clinically actionable genes in WES data. (A) A SNP6 verified *PTEN* deletion is missed by TNR and FACETS due to noisy read depth signal, but caught by `dryclean`. (B) *MET* amplification is missed by TNR and FACETS but caught by `dryclean`. (C) A missed *EGFR* amplification in FACETS and TNR due to an artifactual event in the normal sample which skews the TNR. Both `dryclean` and SNP6 detect this event. For `dryclean` and TNR both raw data and segmentations from CBS are shown.



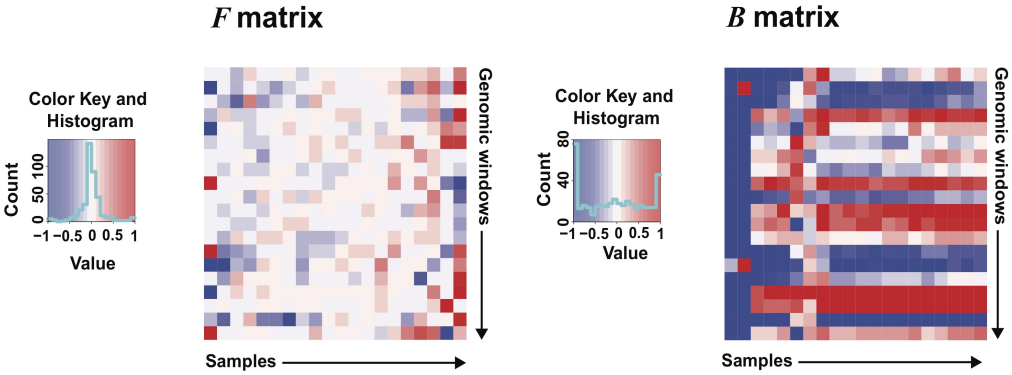
Supplementary Figure S5. Paired normal sample shows a broad and high amplitude EGFR amplification. The two tracks show the raw read counts for tumor (lower track) and paired normal sample (upper track). It is clear from the read counts for these samples that the putative normal sample has 10 times as many reads mapping to *EGFR* locus. This explains the deletion signal seen in TNR and FACETS (Figure S4) even though they use different types of information, both methods ultimately normalize with paired normal sample.



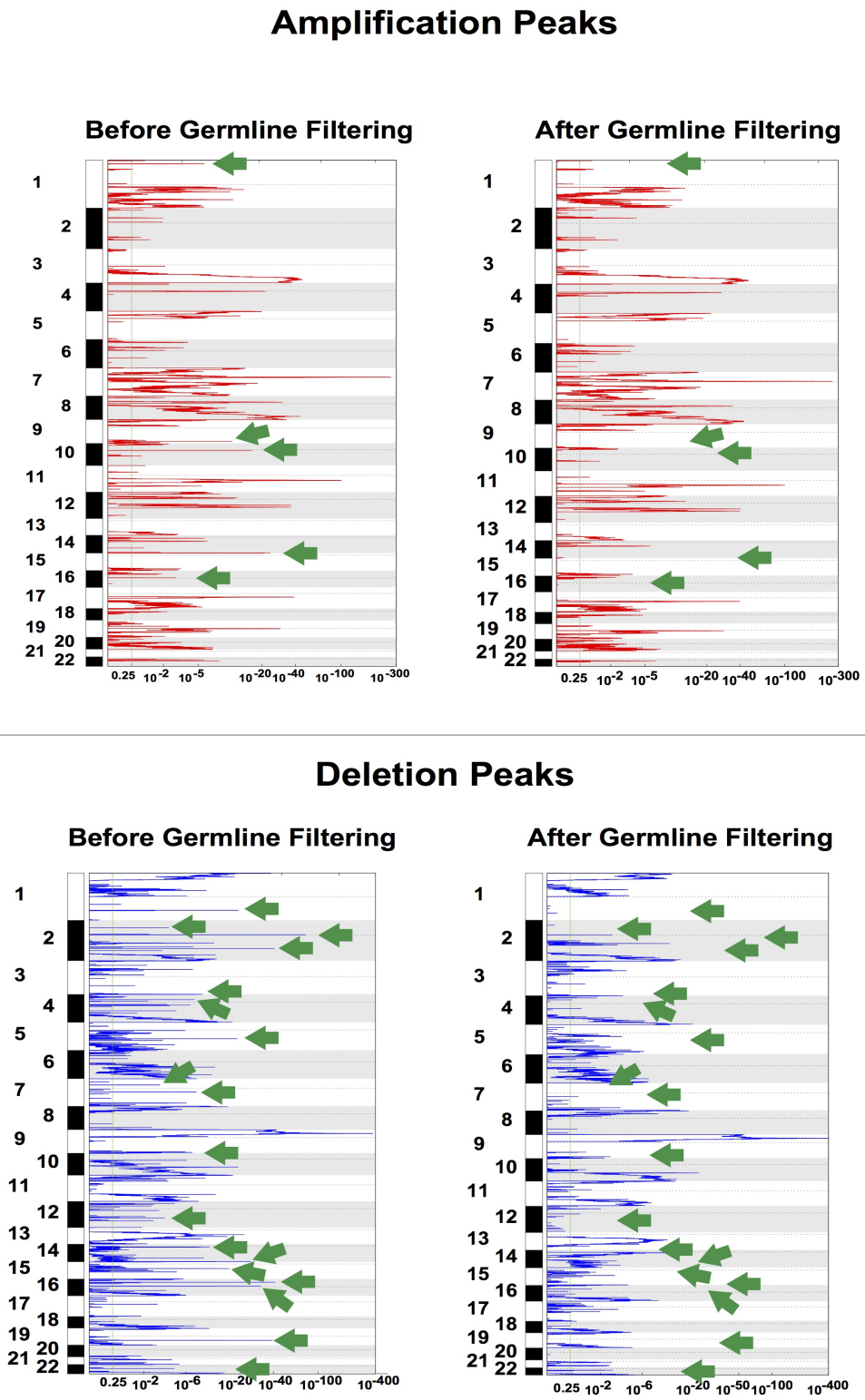
Supplementary Figure S6. Quality of WES read depth data improves with number of normal samples in the PON. We mesure the quality of read depth data in terms of derivative log ratio spread (DLRS). We see reduction in DLRS value as the number of normal samples in the PON increase. Interestingly, there is a significant reduction in noise compared to TNR with a PON composed of as few as 50 samples (Wilcoxon rank sum test, p value $< 2.2 \times 10^{-16}$). When we add germline filter as described in main text, we get the best results. TNR: tumor-normal ratio, 50 PON: dryclean with 50 normal samples in panel of normal (PON). 100 PON: dryclean with 100 normal samples in PON.



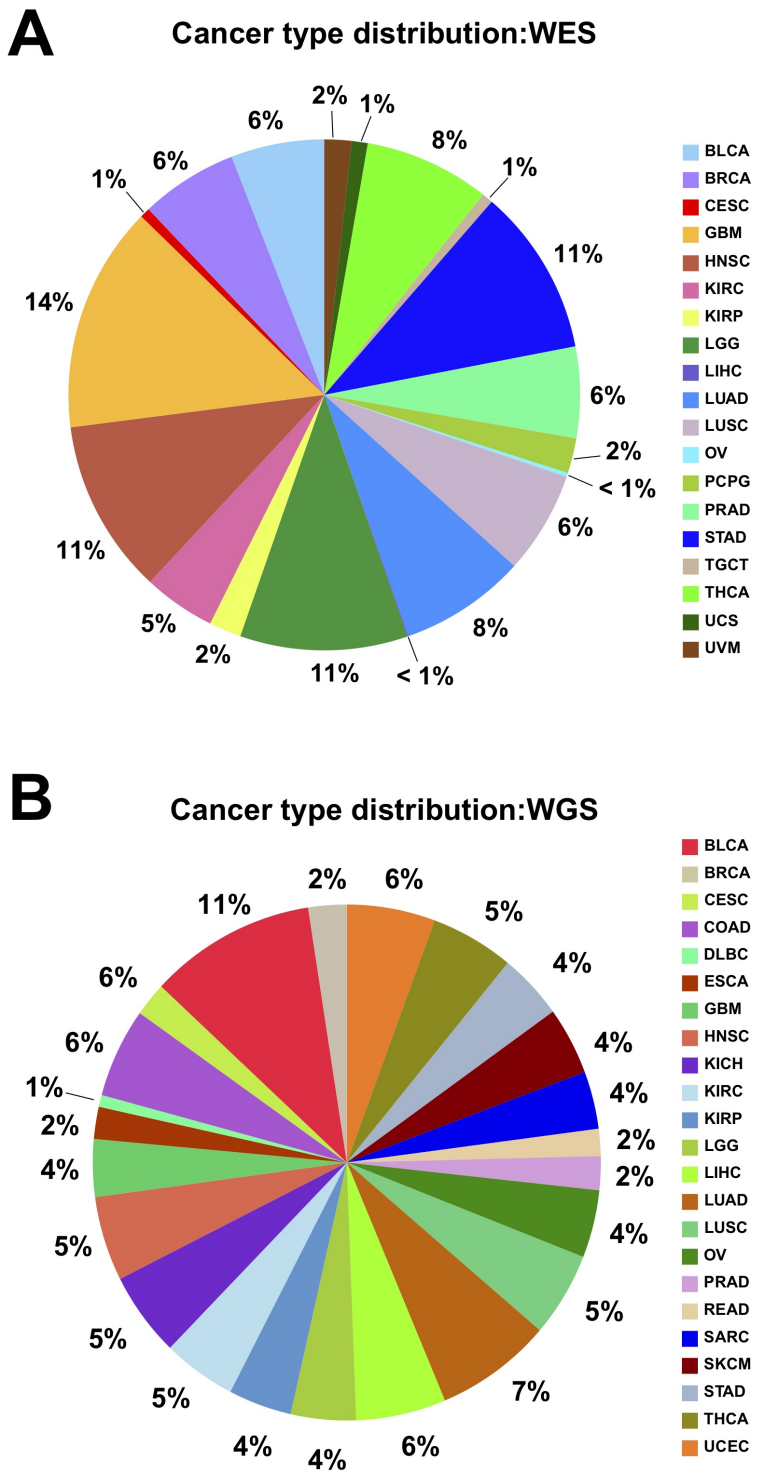
Supplementary Figure S7. Rich structure in the genomic background subspace of WES normal samples. Shows the structure when 200 random WES samples were clustered using the inferred genomic background subspace. A small region of genome (chr. 22) was used to extract the background subspace. `dryclean` uses this approach to make a more informed decision to select normal samples to be included in PON.



Supplementary Figure S8. Visualization of $\mathbf{B}$, $\mathbf{F}$ matrices after decomposition of PON. Each column here is a sample and each row is a genomic window. The values are normalized read counts in log space. The $\mathbf{F}$ matrix is mostly sparse and captures germline events. The $\mathbf{B}$ matrix captures genomic background. Here we see at least three distinct clusters, indicating the heterogeneity of biological and technical noise among samples.



Supplementary Figure S9. Effect of germline filter on the recurrent SCNA output from GISTIC2.0. The amplification and deletion peaks from dryclean before and after germline filter implementation within the algorithm. The green arrows point to the narrow peaks that are putative germline polymorphic events and are removed after application of the CNP filter.



Supplementary Figure S10. Summary of TCGA samples used in analysis. (A) Distribution of 19 tumor types in the WES sample set with 1576 samples. (B) Distribution of 23 tumor types in the WGS sample set with 943 samples

Supplementary Table S1: Raw counts for F1 calculations in Figure 5.

`dryclean` WES						
SCNA	True Positive	False Negative	False Positive	Recall	Precision	F1
EGFR amp in Lung	115	10	4	0.92	0.97	0.94
EGFR amp in GBM	195	7	6	0.97	0.97	0.97
MET amp in KIRC	26	2	0	0.92	0.97	0.94
MET amp in Lung	89	3	7	0.97	0.93	0.95
FGFR1 amp in LUSC	48	0	1	1	0.98	0.99
CDKN2A del in LUSC	158	4	7	0.98	0.96	0.97
PTEN del in PRAD	28	0	2	1	0.93	0.96
MDM2 amp in Lung	64	3	6	0.96	0.91	0.93
ERBB2 amp in BRCA	34	1	0	0.97	1	0.98
Total	731	28	33	0.96	0.96	0.96

`dryclean` Panel						
SCNA	True Positive	False Negative	False Positive	Recall	Precision	F1
EGFR amp in Lung	102	22	5	0.82	0.95	0.88
EGFR amp in GBM	197	5	4	0.98	0.98	0.98
MET amp in KIRC	25	3	1	0.89	0.96	0.92
MET amp in Lung	80	12	8	0.87	0.91	0.89
FGFR1 amp in LUSC	46	1	0	0.98	1	0.99
CDKN2A del in LUSC	144	16	7	0.90	0.95	0.92
PTEN del in PRAD	25	2	1	0.93	0.96	0.94
MDM2 amp in Lung	56	10	4	0.85	0.93	0.89
ERBB2 amp in BRCA	35	0	2	1	0.95	0.97
Total	685	68	31	0.91	0.96	0.93

TNR WES						
SCNA	True Positive	False Negative	False Positive	Recall	Precision	F1
EGFR amp in Lung	108	17	8	0.86	0.93	0.89
EGFR amp in GBM	155	47	7	0.77	0.96	0.85
MET amp in KIRC	24	4	1	0.86	0.96	0.91
MET amp in Lung	80	12	23	0.87	0.78	0.82
FGFR1 amp in LUSC	46	2	6	0.96	0.88	0.92
CDKN2A del in LUSC	154	8	6	0.95	0.96	0.95
PTEN del in PRAD	26	2	2	0.93	0.93	0.93
MDM2 amp in Lung	60	7	14	0.90	0.81	0.85
ERBB2 amp in BRCA	32	3	2	0.91	0.94	0.92
Total	661	98	68	0.87	0.91	0.89

TNR Panel						
SCNA	True Positive	False Negative	False Positive	Recall	Precision	F1
EGFR amp in Lung	101	23	6	0.81	0.94	0.87
EGFR amp in GBM	151	51	6	0.75	0.96	0.84
MET amp in KIRC	24	4	3	0.86	0.89	0.87
MET amp in Lung	73	19	14	0.79	0.84	0.81
FGFR1 amp in LUSC	45	2	7	0.96	0.87	0.91
CDKN2A del in LUSC	142	18	15	0.89	0.90	0.89
PTEN del in PRAD	25	2	2	0.93	0.93	0.93
MDM2 amp in Lung	53	13	7	0.80	0.88	0.84
ERBB2 amp in BRCA	33	2	3	0.94	0.92	0.93
Total	623	130	60	0.83	0.91	0.87

FACETS WES						
SCNA	True Positive	False Negative	False Positive	Recall	Precision	F1
EGFR amp in Lung	114	11	5	0.91	0.96	0.93
EGFR amp in GBM	83	4	0	0.95	1.00	0.97
MET amp in KIRC	23	2	0	0.92	1.00	0.96
MET amp in Lung	82	10	7	0.89	0.92	0.90
FGFR1 amp in LUSC	39	9	4	0.81	0.91	0.86
CDKN2A del in LUSC	146	16	5	0.90	0.97	0.93
PTEN del in PRAD	19	9	1	0.68	0.95	0.79
MDM2 amp in Lung	58	9	3	0.87	0.95	0.91
ERBB2 amp in BRCA	34	1	2	0.97	0.94	0.95
Total	575	69	27	0.89	0.96	0.92

FACETS Panel						
SCNA	True Positive	False Negative	False Positive	Recall	Precision	F1
EGFR amp in Lung	92	32	14	0.74	0.87	0.80
EGFR amp in GBM	117	6	6	0.95	0.95	0.95
MET amp in KIRC	23	2	3	0.92	0.88	0.90
MET amp in Lung	78	14	21	0.85	0.79	0.82
FGFR1 amp in LUSC	28	19	12	0.60	0.70	0.65
CDKN2A del in LUSC	109	51	15	0.68	0.88	0.77
PTEN del in PRAD	12	15	4	0.44	0.75	0.55
MDM2 amp in Lung	47	19	10	0.71	0.82	0.76
ERBB2 amp in BRCA	27	8	4	0.77	0.87	0.82
Total	510	164	86	0.76	0.86	0.81