Comparing DNA double strand break methods and analyses

Linhui Huang (Malinda)[1] and Maureen McKeague[2,3]

[1]School of Computer Science, Faculty of Science, 3480 Rue University, Montreal, QC H3A 2A7, Canada

[2]Pharmacology and Therapeutics, Faculty of Medicine, McGill University, Montreal, QC H3G 1Y6, Canada

[3]Department of Chemistry, Faculty of Science, McGill University, Montreal, QC H3A 0B8, Canada

Correspondence: linhui.huang@mail.mcgill.ca; Maureen.mckeague@mcgill.ca

Introduction

DNA damages in cells are constantly induced by endogenous and exogeneous agents like oxidative stress, chemotherapeutic drugs, UV radiation, environmental pollution, etc. Types of DNA damage include 8-OxodG, platinated crosslinks, abasic sites, single-strand breaks, double-strand breaks, etc. (Mingard et al.). Among those, double-strand breaks (DSB) are relatively more difficult to repair because neither strand is intact to act as a template for template-guided DNA repair mechanisms such as base excision repair and nucleotide excision repair (Dexheimer, 2012). Failure in repairing the damage might lead to mutagenesis or cell death.

In this paper, we compare and analyze the sequencing data obtained using four genome-wide double-strand break mapping techniques: BLESS (Crosetto et al.), DSBCapture (Lensing et al.), BLISS (Yan et al.), and sBLISS (Bouwman et al.) (Table 1). These four techniques all use different adapters to capture DSBs in human cells *in situ*, making their resulting next-generation sequencing data comparable after a similar statistical analysis process. BLESS (direct *in situ* breaks labeling, enrichment on streptavidin and next-generation sequencing) and DSBCapture use adapters to directly tag DSBs and achieve single-nucleotide resolution tagging of DSBs. On the other hand, BLISS (Breaks Labeling *In Situ* and Sequencing) and sBLISS (in-suspension BLISS) blunt DNA before attaching adapters, making them unable give exact DSB positions. What is unique about the adapters of BLISS and sBLISS is that they contain unique molecular identifiers (UMIs) that allow for quantification of DSBs at each break position.

The goal of this project is to investigate the sequencing data from the four methods and compare break sequence contexts, sensitivity of selected cancer genes and relative number of

breaks. Genome-wide sequencing and mapping of DNA double-strand breaks is essential for elucidating mechanisms of disease initiation and evaluating the efficacy of chemotherapeutic drugs.

Table 1: Genome-wide DNA double-strand break mapping techniques mentioned in this paper

Mapping Method	**DNA Source**	**Drug Treatment**	**Ref.**
BLESS	Human HeLa cells	Untreated vs. 0.4uM aphidicolin treatment for 18h	Crosetto et al.
DSBCapture	Human NHEK cells	Untreated	Lensing et al.
BLISS	Human U2OS cells	Untreated vs. 10uM or 150uM etoposide treatment	Yan et al.
sBLISS	Human TK6 cells	Untreated vs. 30uM etoposide treatment	Bouwman et al.

Methods

2.1. Data

At the time of doing this project, BLESS, DSBCapture, BLISS, and sBLISS are the papers of sequencing DNA double-strand breaks with available human cell data. DNA sequencing data from the four papers are downloaded from NCBI Sequence Read Archive using the accession codes SRP018506 (BLESS) and SRP099132 (BLISS) and from Gene Expression Omnibus using the accession codes GSE78172 (DSBCapture) and GSE145598 (sBLISS).

2.2. Processing raw data

Umi_tools, cutadapt, and samtools are used to remove the UMI and adapters on the DNA sequencing data. Bowtie is used to align the sequences to the human reference genome (hg19 for BLESS and DSBCapture and hg38 for BLISS and sBLISS). Alignments in the blacklist regions of the human genome (obtained from https://sites.google.com/site/anshulkundaje/projects/blacklists) are removed.

2.3. Individual analyses

To obtain sequence motif around the DSBs, sequences 5 bp upstream and 5 bp downstream of each DSB is recorded. The frequency of each base is calculated and plotted in Figure 1.

For experiments with both drug-treated and untreated samples, hypergeometric p-values are calculated for each 10,000-bp window on the genome to identify important windows of high DSB abundance. For experiments with only untreated samples, Poisson p-values of the same windows are calculated.

For a list of cancer genes identified by cancer gene census (Cosmic, 2017), the hypergeometric or Poisson p-value of each cancer gene window, defined as from the transcription start site to the transcription end site, is also calculated. The $-\log_{10}(\text{p\_value})$ are plotted in Figure 2 to rank the relative sensitivity of the cancer genes.

The numbers of DSBs under different drug concentrations of the four methods are plotted in Figure 3. Note that all four methods use different types of cells, and only BLISS and sBLISS apply the same drug, etoposide.

The data analysis code is available at https://github.com/MalindaH/DNA-Break-Analysis.

Results

For BLESS, there is no difference between the break sequence motifs of the treated and untreated samples. The most common sequence motif of BLESS has A one position upstream, G one position downstream, and A two positions downstream (Figure 1A). For DSBCapture, the sequence context shows that the frequencies of bases at most positions match the GC content of the reference human genome of about 40% (Piovesan et al.), except that the first and third positions downstream of the DSB have a slightly higher GC frequency of about 50%.

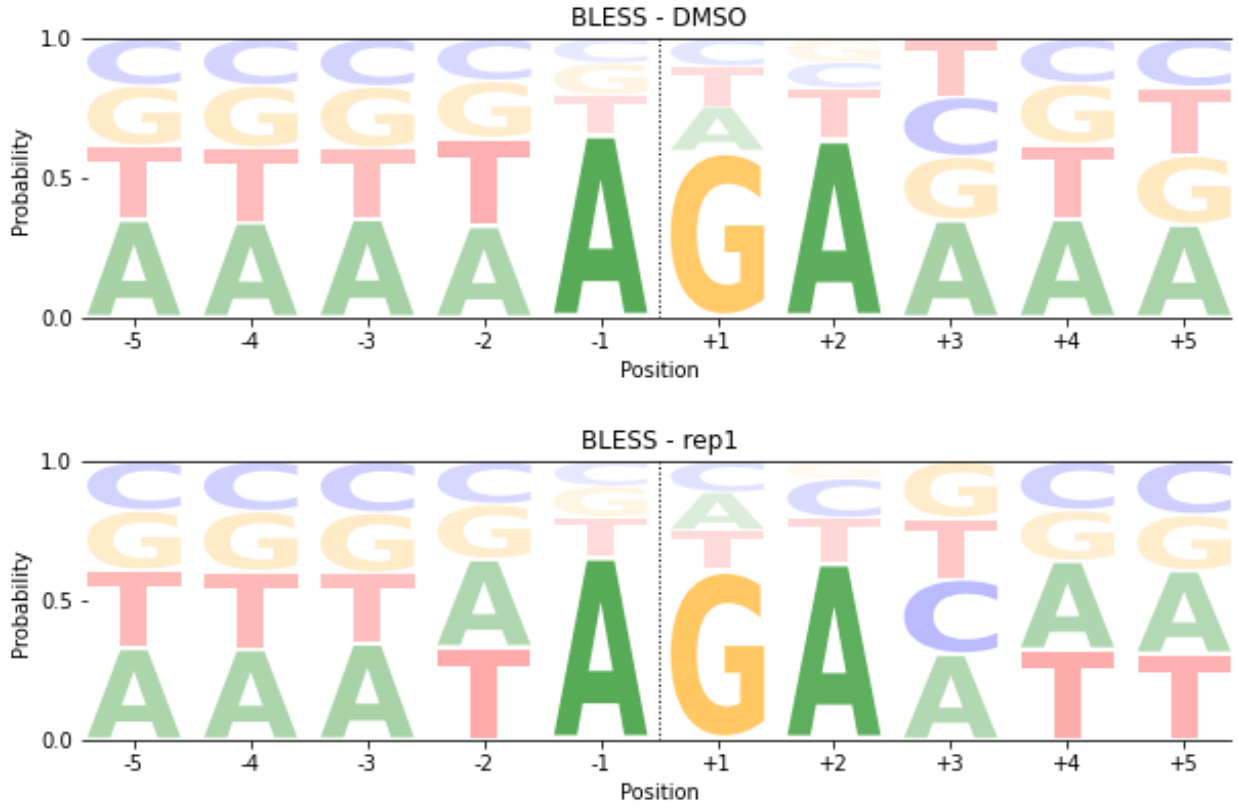
Figure 2 shows the top 20 genes cancer genes ranked by relative drug sensitivity according to break frequencies upon drug treatments of BLISS, sBLISS, and BLESS. This cancer gene sensitivity analysis was also done by BLESS and we reproduce this analysis with the other datasets as well. Some genes that are commonly at the top of the lists are Oligodendrocyte Transcription Factor 2 (OLIG2), HIST1H3B, and C-X-C Motif Chemokine Receptor 4 (CXCR4). OLIG2 is commonly associated with T-Cell Acute Lymphoblastic Leukemia and affects protein homodimerization activity, transcription factor activity, and RNA polymerase II distal enhancer sequence-specific binding. HIST1H3B makes the H3 Clustered Histone 2

protein, which is a core component of nucleosomes, and is associated with hematologic cancer and brain stem cancer. CXCR4 is associated with Whin syndrome and affects G protein-coupled receptor activity and ubiquitin ligase binding.

In addition, the break abundances in relation with drug treatment concentrations of the four methods are shown in Figure 3. The higher the drug concentration, the more DNA double-strand breaks are induced.

We also analyzed increases in DSB distributions caused by drug treatments over each chromosome as well as over individual protein-coding genes and found interesting trends among a group of genes, an increase in DSB counts around transcription start sites and a drop in DSB counts around transcription termination sites. More details are in supporting information on our website (https://github.com/MalindaH/DNA-Break-Analysis).

Figure 1: DSB sequence motifs of BLESS and DSBCapture. X-axis: base pair positions relative to the DSB. Negative numbers indicate upstream positions and positive numbers indicate downstream positions.



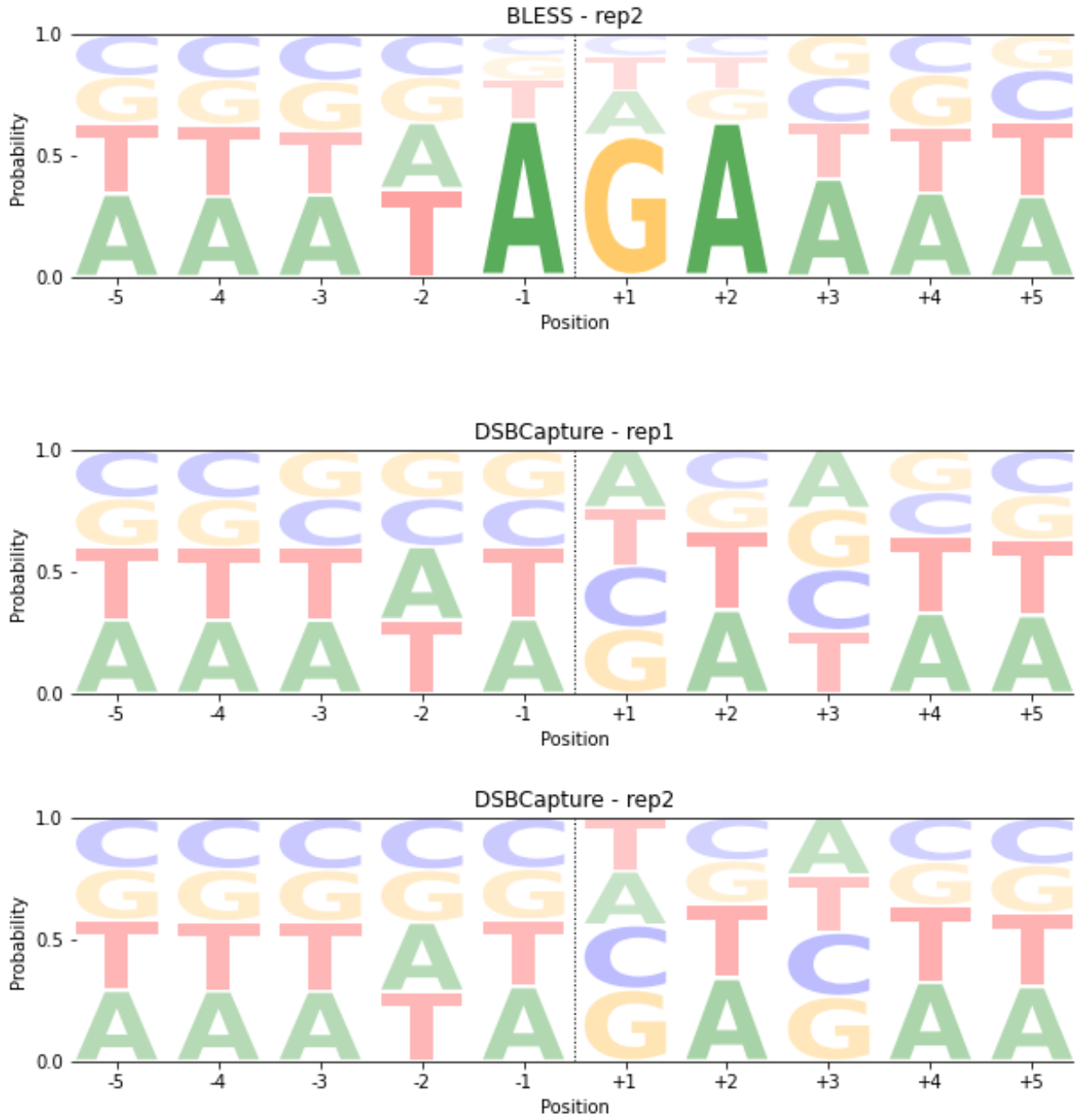


Figure 2: Top 20 cancer genes most sensitive to drug treatments of BLISS, sBLISS, and BLESS. The genes are plotted in decreasing order of average relative break abundance among the three methods.

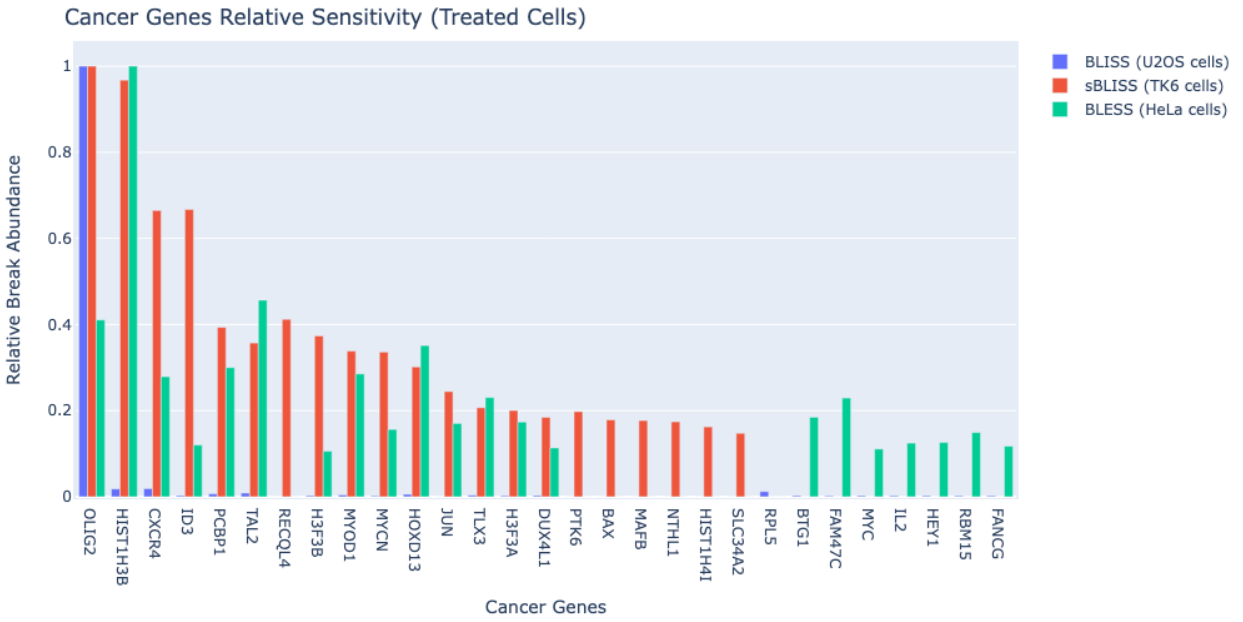
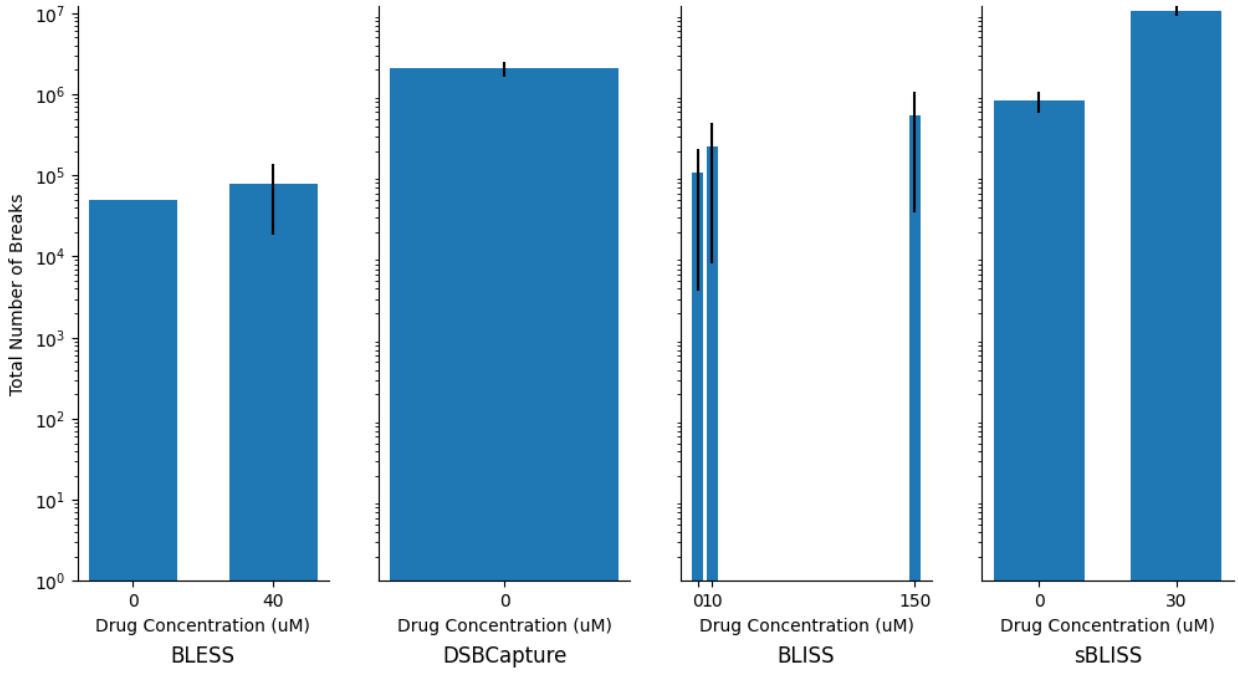


Figure 3: Total number of DSBs upon different drug treatment concentrations of BLESS, DSBCapture, BLISS, and sBLISS.



Discussion

Click-Code-Seq (Wu et al., 2018) investigated sequence contexts of 8-oxoG damage. However, no paper has ever looked at break sequence contexts of DSBs, so we analyzed

sequence motifs of DSB based on the four datasets. The DSB sequence motif analyses (Figure 1a) of BLESS shows that human HeLa cells have a characteristic DSB distribution at sequences with A one position upstream, G one position downstream, and A two positions downstream. The fact that this sequence motif is not significantly different between untreated and treated samples suggests that this distribution pattern corresponds with the cell type and not the type of drug treatment. The sequence motif of untreated samples from DSBCapture (Figure 1b) is also observed to have slightly higher GC frequency of about 50% at the first and third downstream positions. This, again, suggests that different cell types inherently have different DSB sequence motifs. This finding may be useful to aid in designing drugs that target DNA sequences in specific cell types. Also, the adapters used in BLISS and sBLISS contain UMIs and do not mark the exact break sites, so no sequence motif can be concluded from BLISS and sBLISS datasets.

Among the list of cancer genes analyzed, we identify OLIG2, HIST1H3B, and CXCR4 as the most sensitive to drugs, which are commonly found on BLISS, sBLISS, and BLESS datasets (Figure 2). This cancer gene sensitivity analysis can be performed on other genes of interest as well to elucidate gene's sensitivity to specific drugs which may infer the possibility of using the drugs as chemotherapeutics to target specific genes.

The remaining analysis show that the higher the drug concentration, the higher the DSB break abundance (Figure 3). Also, note that the absolute number of breaks between these methods are orders of magnitude different: even between BLISS and sBLISS where the same drug treatment is applied to human U2OS and TK6 cells, the break numbers are 10x to 100x higher with sBLISS. This indicates a different susceptibility of cell types to different drugs.

Our findings pose important insights into designing novel chemotherapeutics. However, the four methods that we analyze involve different drug treatments of different cell types, it is hard to make quantitative conclusions about the observed trends. In a future study, specific drug treatments and cell lines should be selected to make direct comparisons of the cells' susceptibility to the drugs.

Conclusion

For the first time we compare four methods, namely BLESS, DSBCapture, BLISS, and sBLISS, on the distribution of DSBs upon drug treatments. We have found that different cell types have specific DSB sequence motifs. We have identified some common cancer genes from

BLESS, BLISS, and sBLISS that are more sensitive to drug treatments than others, and confirmed that the number of DSBs positively correlates with the drug treatment concentration. These analyses can be used in future studies to confirm drug effectiveness and target accuracy on cell lines for developing novel chemotherapeutics.

References

Bouwman, B. A. M., Agostini, F., Garnerone, S., Petrosino, G., Gothe, H. J., Sayols, S., Moor, A. E., Itzkovitz, S., Bienko, M., Roukos, V., & Crosetto, N. (2020). Genome-wide detection of DNA double-strand breaks by in-suspension BLISS. *Nature Protocols*, *15*(12), 3894–3941. https://doi.org/10.1038/s41596-020-0397-2

Cosmic. (2017). *Cancer Gene Census*. Cancer.sanger.ac.uk. https://cancer.sanger.ac.uk/census

Crosetto, N., Mitra, A., Silva, M. J., Bienko, M., Dojer, N., Wang, Q., Karaca, E., Chiarle, R., Skrzypczak, M., Ginalski, K., Pasero, P., Rowicka, M., & Dikic, I. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nature Methods*, *10*(4), 361–365. https://doi.org/10.1038/nmeth.2408

Dexheimer, T. S. (2012). DNA Repair Pathways and Mechanisms. *DNA Repair of Cancer Stem Cells*, 19–32. https://doi.org/10.1007/978-94-007-4590-2_2

Lensing, S. V., Marsico, G., Hänsel-Hertsch, R., Lam, E. Y., Tannahill, D., & Balasubramanian, S. (2016). DSBCapture: in situ capture and sequencing of DNA breaks. *Nature Methods*, *13*(10), 855–857. https://doi.org/10.1038/nmeth.3960

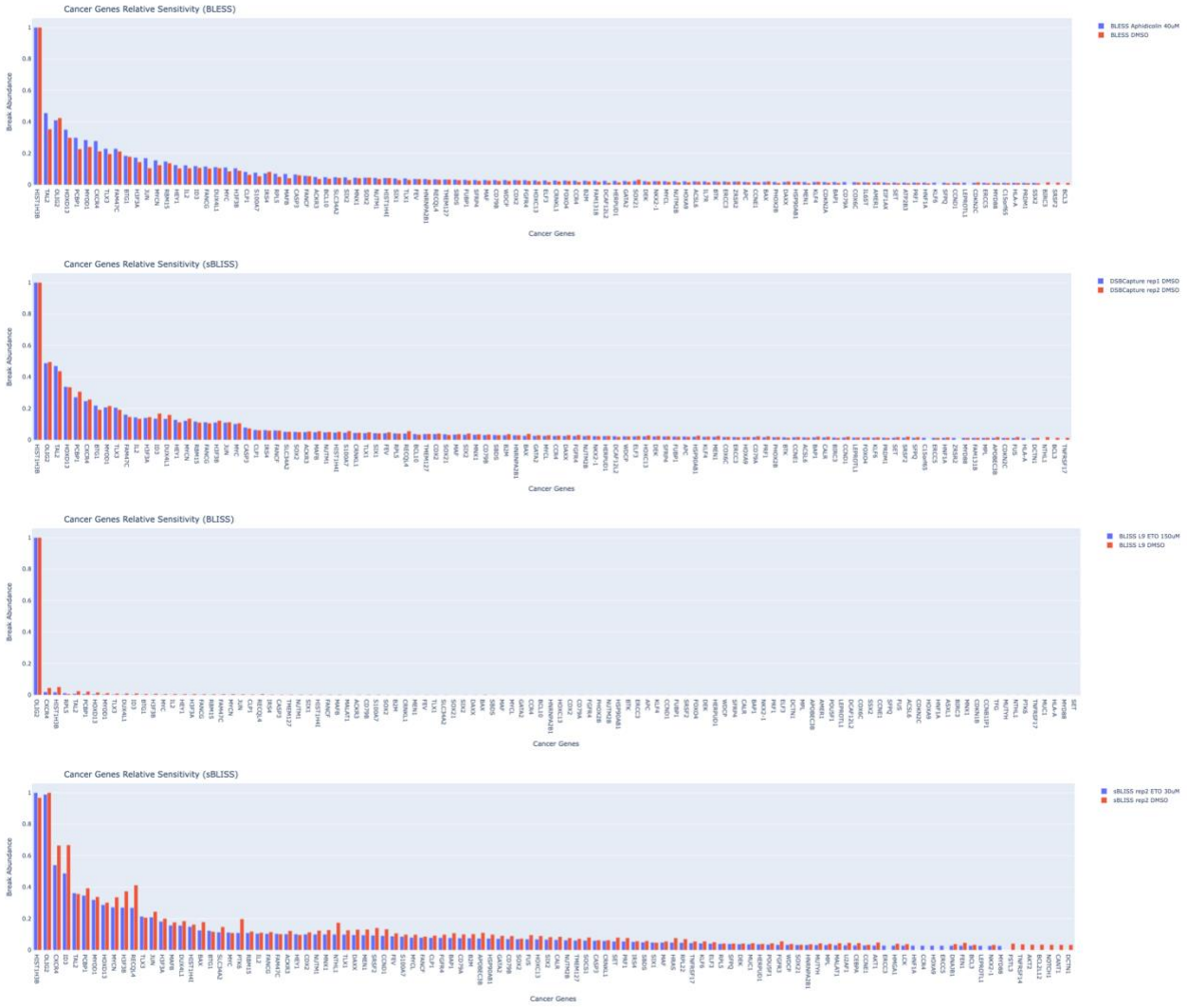
Mingard, C., Wu, J., McKeague, M., & J. Sturla, S. (2020). Next-generation DNA damage sequencing. *Chemical Society Reviews*, *49*(20), 7354–7377. https://doi.org/10.1039/D0CS00647E

Piovesan, A., Pelleri, M. C., Antonaros, F., Strippoli, P., Caracausi, M., & Vitale, L. (2019). On the length, weight and GC content of the human genome. *BMC Research Notes*, *12*(1). https://doi.org/10.1186/s13104-019-4137-z

Wu, J., McKeague, M., & Sturla, S. J. (2018). Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *Journal of the American Chemical Society*, *140*(31), 9783–9787. https://doi.org/10.1021/jacs.8b03715

Yan, W. X., Mirzazadeh, R., Garnerone, S., Scott, D., Schneider, M. W., Kallas, T., Custodio, J., Wernersson, E., Li, Y., Gao, L., Federova, Y., Zetsche, B., Zhang, F., Bienko, M., & Crosetto, N. (2017). BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications*, *8*(1). https://doi.org/10.1038/ncomms15058

Appendix


Cancer Genes Relative Sensitivity (BLESS)
Break Abundance
Cancer Genes
BLESS Aphidicolin 40uM
BLESS DMSO
Cancer Genes Relative Sensitivity (sBLISS)
Break Abundance
Cancer Genes
DSBCapture rep1 DMSO
DSBCapture rep2 DMSO
Cancer Genes Relative Sensitivity (BLISS)
Break Abundance
Cancer Genes
BLISS L9 ETO 150uM
BLISS L9 DMSO
Cancer Genes Relative Sensitivity (sBLISS)
Break Abundance
Cancer Genes
sBLISS rep2 ETO 30uM
sBLISS rep2 DMSO