Referee comments

We would like to thank the reviewers for their insightful comments that have allowed us to improve our work in each iteration. Our point-by-point answers follows, the original comments of the reviewers are in black, our responses are in blue.

Referee: 2

In this version of the manuscript, the major criticisms from the previous version have been addressed satisfactorily. The authors clearly distinguish between OneD and OneD+CN and show its virtues with a separate experiment. The speed benchmark for OneD is convincing and a positive feature which should increase the impact of this method. The method is explained in more detail and more clearly in this iteration. It is acceptable to me that Servant et al. preprint is not included for comparison, since it is not peer-reviewed yet. I believe my language made it clear that the potential reason for its exclusion is most likely the fact that it is a very recent reprint, not the authors' lack of scientific rigor. It is reiterated here for clarification.

I have some major suggestions that can improve the claims in the paper, followed by some minor edits to text.

Major points

Some abbreviations are introduced before their definition. Unless I am missing it, vanilla normalization (which I guess is running a single round of ICE) is not defined. Please make sure all abbreviation are explained at their first appearance in the text.

This has been fixed.

Is there a reason HiCNorm is referred to as LGF? HiCNorm is known by its name in the field, so I suggest using 'HiCNorm' to refer to it.

The reason was that LGF is the principle behind method (Local Genomic Feature) rather than a particular implementation. However, we agree that the norm should prevail, so we have replaced LGF by HiCNorm in the text and figures.

Overdispersion is not defined. While it is a widely known concept for computational genomicists, authors should define and cite some papers to explain its importance. Carty et al. in their HiCDC paper (doi:10.1038/ncomms15454) talk about its relevance to Hi-C. I believe the original and well known DESeq paper introduces it and the application negative binomial distribution to adrres it for sequencing data.

We have addressed this issue in the revised text, including the suggested citations.

Fig1B is very useful and clearly demonstrates the difference between OneD and OneD+CN. If array data for CNV is available, it should be plotted alongside this data to conclusively show the

CNV profile tracks normalization factors. Otherwise, another dataset with array data can be used.

Available CN array data did not cover the region shown in Figure 1B. We had to change the figure, and we now show a region of chromosome 9 with karyotypic imbalance for which array data is available.

In the ‘aberrant karyotype’ experiment, can the authors offer any insight as to why calCB performs better than OneD+CN whereas it is vice versa for ‘copy number correction’ experiment. This trend seems more obvious for K562. The ‘copy number correction’ experiment clearly demonstrates the virtue of OneD+CN and its improved performance. Is it possible K562 contains many rearrangements that further complicate this task?

caICB takes into account possible whole chromosome amplifications / deletions whereas OneD+CN allows for local copy number variations. If the purpose is to remove the effect of copy number, then OneD+CN outperforms caICB because it allows for more possible genotypes. On the other hand, if the purpose is to improve the reproducibility among samples of the same type, correcting the copy number removes part of the signal, and therefore blurs the distinction between experiments.

The information has been added to the discussion.

For the Capture Hi-C section, reproducibility score metric is not an appropriate choice for assessing reproducibility. Hi-C Spector is tailored for Hi-C data and its behaviour on Capture Hi-C data is not tested. It is the same for SCC metric. It would be more appropriate to use correlation coefficient for such datasets.

Here we are not using promoter capture Hi-C. Instead, we capture contiguous fragments so the resulting matrices share the same structure as Hi-C data. Hi-C Spector is tailored to Hi-C data, but for the sake of consistency with the rest of the results, we chose to keep the Hi-C Spector reproducibility score in the main Figure 5. We have added the Spearman and Pearson correlation in Supplementary Figure 14.

I am likely missing the URL but is OneD software available? If I have overlooked the link, please ignore this point.

OneD is of course available. We mentioned the URL in the last paragraph of the “Data sources” subsection (https://github.com/qenvio/dryhic). We have also added the URL in the abstract.

The Hi-C datasets authors gathered for this benchmark can serve the field for future related benchmarks. I realize the authors use existing datasets and clearly display the sources of each experiment. However, if this data is bundled together and available for use in a single location, this would benefit the field. I would like to encourage them to create such a resource, thought it is at their discretion.

We thank the reviewer for such a good suggestion. As the raw data are available in their corresponding repositories (see Supplementary Table 1) we have set a dedicated server with the matrices at 100 Kbp resolution. Its location can be found in the front page of the method URL (https://github.com/qenvio/dryhic).

This information has been added to the abstract.

Minor points

What is the resolution of data used to make Fig1A? Mention in legend.

The resolution has been specified.

Per line 49 in intro, I am not sure the reason ICE is popular is its speed but more its simplicity. This claim can be reconsidered.

We have changed the sentence accordingly.

Page 11, Line 31: “The values are stored in the cells of the contact matrix (xij), quantifying the interaction between the two loci at positions i and j.” change to “…quantifying the interaction Frequency…”

This has been done.

Page 12, Line 11: There is a typo: “karyotyp”.

The typo has been corrected.

What are CRG and UCSC?

CRG stands for Centre for Genomic Regulation. UCSC was a typo and it has been amended to UCSD (University of California, San Diego). Both acronyms were defined in Supplementary Table 1 legend, but they have been redefined at the first occurrence in the main text.

Simulated datasets can be visualized in the supplement to visually convince readers.

We have added a figure (Supplementary Figure 13) with a representative example of the simulated datasets.

“For instance, a duplicated region in a diploid genome will show on average a 50% increase in the number of contacts.” This seems intuitively correct but is there any support for this number?

Hi-C data is quantitative. This is best demonstrated by the fact that duplications can be detected and resolved from Hi-C data to reassemble genomes (see for instance GRAAL, Marie-Nelly *et* *al*., 2014; Nature Communications doi:10.1038/ncomms6695). The fact that the Hi-C signal is proportional to the amount of starting material is globally accepted and does not require further justification.

Page 14, Line 51: “Even though the variability is high in this case, this suggests that the karyotypes of these cells have diverged, so that the copy number correction performed by OneD increased the reproducibility between samples from the same cell type.” Do the authors intend to say OneD+CN here? I have the same question for OneD statement in Page 15, Line 27.

Yes. This has been changed.

Page 16, line 37: “..downsampling the contact matrices to mimic lower resolutions..” Change resolution to sequencing depth or coverage. Resolution refers to bin-size in this manuscript.

This has been corrected.

Referee: 1

Comments for the Author
I am satisfied with the response from the authors. They have addressed all my concerns.