Brief Report

Benchmarking State-of-the-Art Approaches for Norovirus Genome Assembly in Metagenome Sample

Dmitry Meleshko [1] and Anton Korobeynikov [1,2,*]

1   Center for Algorithmic Biotechnology, St. Petersburg State University, 7/9 Universitetskaya Emb., 199004 St. Petersburg, Russia; d.meleshko@spbu.ru
2   Department of Statistical Modelling, St. Petersburg State University, Universitetskiy 28, 198504 St. Petersburg, Russia
*   Correspondence: a.korobeynikov@spbu.ru

Simple Summary: Proper use of state-of-the-art tools allows for complete recovery of viral genomes from a metagenome sample.

Abstract: A recently published article in *BMC Genomics* by Fuentes-Trillo et al. contains a comparison of assembly approaches of several noroviral samples via different tools and preprocessing strategies. It turned out that the study used outdated versions of tools as well as tools that were not designed for the viral assembly task. In order to improve the suboptimal assemblies, authors suggested different sophisticated preprocessing strategies that seem to make only minor contributions to the results. We have reproduced the analysis using state-of-the-art tools designed for viral assembly, and we demonstrate that tools from the SPAdes toolkit (RNAVIRALSPADES and CORONASPADES) allow one to assemble the samples from the original study into a single contig without any additional preprocessing.

Keywords: benchmarking; viral assembly; noroviruses

1. Introduction

In recent years, there has been a significant surge of interest in the field of viral discovery, with numerous large-scale studies being conducted [1–3]. Many research groups have shifted their focus to viral studies, therefore exploding the amount of papers devoted to the subject. Despite the considerable attention given to viral genome assembly, there is currently no established state-of-the-art method for this task. The diversity of viral genomes and corresponding sequencing data presents a challenge, as assembly approaches suitable for dsDNA bacteriophages cannot be directly applied to RNA viruses or retroviruses without appropriate modifications. As a result, some papers are naturally devoted to benchmarking viral assembly approaches for different kinds of input data. However, proper benchmarking design (cf. [4–7] as an example of well-designed studies for non-viral data) and interpretation of the obtained results is a non-trivial task, and subtle flaws made here could easily lead to misleading or controversial outcomes.

An article discussing different approaches for norovirus genome assembly was recently published in *BMC Genomics* [1]. Using the approaches presented in this article, we want to highlight some common benchmarking problems that might result in misleading conclusions. By highlighting these problems, we hope to emphasize the importance of avoiding such pitfalls in future studies.

The article was submitted and subsequently published in 2021; however, the tools utilized in the study were notably outdated. Specifically, a comparison of genome assemblers METASPADES v.3.11.1 [8] and MEGAHIT v.1.1.3 [9] was conducted. Notably, METASPADES 3.11.1 was released in March 2018, while the current version of SPAdes is 3.15.5, and the SPAdes team makes multiple releases each year. SPAdes 3.15.0 was released

on January 2021 and was available to use in this study. Similarly, MEGAHIT v.1.1.3 was released in March 2018, while the most recent version, 1.2.9, was released in October 2019. In scientific research, the use of legacy tool versions can be justified based on reasons such as reproducibility, reliability, and compatibility between different tools and pipelines. Moreover, the extended duration of the scientific publishing process can make it impractical to update the tool versions used at every revision, which can take months in some cases. However, in the case of this article, the utilization of excessively outdated tool versions raises concerns because of the substantial time gap between the release of the tool versions employed and the time of publishing. This time gap undermines the applicability of any of the aforementioned reasons for legacy software usage as a valid justification. Therefore, it is important to critically evaluate the impact of using outdated tools alongside claims of superior performance, as this approach may lead to potential misinterpretations and fail to accurately represent the current state of the art. This concern is particularly pertinent in benchmarking studies, where the comparison of tools constitutes the primary outcome of the research, compared to papers that offer novel biological insights.

Moreover, the selection of tools for assembly benchmarking raises questions, and no further justification is provided for this choice. It is worth noting that both METASPADES and MEGAHIT are metagenomic assemblers, and while they have shown their effectiveness in non-metagenomic settings, including viral assemblies [10,11], the absence from the study of specialized assemblers designed for RNA and RNA viral data is concerning and may indicate an improper benchmark design. Given that noroviruses are RNA viruses, it is surprising that no dedicated transcriptome or viral assembler was evaluated. Long before the time of the study, there were several prominent assemblers readily available for consideration, such as TRINITY [12], RNASPADES [13] (included in the used SPADES v.3.11.1 package), Savage [14], and IVA [15], among others. Moreover, a dedicated RNA viral assembler, RNAVIRALSPADES [16], had been developed, with the preprint and the tool itself being publicly available since the summer of 2020.

Furthermore, it is worth noting that the authors of the study did not adhere to the recommended best practices for genome assembly as outlined in the SPADES and MEGAHIT manuals. One of the important guidelines is to select a *k*-mer length that is approximately half the length of the input sequencing reads. However, the authors reported a mean read length of 100 bp and tested various *k*-mer lengths, including extremely small values such as 21 and exceptionally long values such as 99 and 119. It is important to highlight that reporting these exceedingly long *k*-mer values can create a false impression that tools such as MEGAHIT and METASPADES are unable to handle such *k*-mer lengths effectively. In reality, these values are not suited for the given read length, since it is impossible to build a meaningful de Bruijn graph with such parameters from these input data.

Finally, the presentation of several statistics that the authors used to show the results could be improved. The study involved the assembly of a total of eight datasets, and the authors reported mean assembly statistics values across these datasets. From experience, these values might vary a lot between assembly runs. Relying solely on mean N50, contig size, and the number of contigs provides limited information about the overall assembly results. Furthermore, different assemblers may employ distinct heuristics for contig filtering, making the direct comparison of these values for benchmarked approaches challenging, especially in a metagenomic context.

To provide a more comprehensive assessment, it would be beneficial to reformat the resulting tables, allowing readers to compare various statistics across multiple datasets and evaluate their variability. While the authors partially addressed this issue by aligning contigs using BLAST and reporting mean values, it is crucial to consider the limited sample size of only eight datasets. By presenting the statistics in a format that facilitates comparison across datasets, readers can gain a better understanding of the performance variation and make more informed assessments of the benchmarked approaches.

To showcase the effect of tool choice, we reproduced the analysis from the mentioned study using the state-of-the-art version of tools and showed that the majority of them do

correctly assemble the noroviral datasets in question into a single contig without requiring additional complex clustering procedures that were previously suggested in the article as a means to address assembly result deficiencies. It is worth noting that the benchmarks performed by the original authors are readily reproducible, as the necessary data and the versions of the tools used were explicitly stated.

2. Methods

The raw data obtained in the original study [1] consist of eight fecal samples collected from patients affected by acute non-bacterial gastroenteritis. The samples were subjected to polyA enrichment using TruSeq RNA Sample Prep Kit v2, followed by Illumina HiSeq 2000 sequencing. For each sample, 15–69 million paired reads of 100 bp were generated. Input data were downloaded from NCBI SRA.

For viral assembly, benchmarking input reads were trimmed using BBDuk with `trimpolya=15 qtrim=rl trimq=10` parameters, as in [16], and assembled using CORONASPADES 3.15.4, RNAVIRALSPADES 3.15.4, RNASPADES 3.15.4, TRINITY 2.13.2, and MEGAHIT 1.2.9, with the default parameters including k-mer sizes. Exact command lines are available in the Supplementary Text. The assembly was performed on a local server with Intel Xeon E7-4800 v2 CPU using 16 computational threads.

CORONASPADES requires a curated set of HMM profiles that matches a particular viral family. While the sets of coronaviral, HIV, and influenza HMMs are readily available from the CORONASPADES website, most of the viral families (including noroviruses) were left uncovered.

Since noroviruses are small ($\approx$ 7500 nt) and have quite simple internal genome organization comprising 3 almost non-overlapping ORFs that encode several polyproteins [17], we prepared a set of norovirus HMM models for CORONASPADES, extracting the models from the RVDB-prot-HMM database [18] with "norovirus" in the keywords. In total, 13 HMM models were obtained, representing different protein families of norovirus genomes for sequences deposited in GenBank.

Scaffolds produced by each assembler were aligned to corresponding references using QUAST v 5.0.2 [19].

3. Results

The quality assessment results obtained are summarized in Table 1. They clearly show that RNAVIRALSPADES and CORONASPADES (the version of RNAVIRALSPADES that could use profile HMM models to guide an assembly) are better suited for the assembly of these data than RNASPADES, as well as METASPADES and MEGAHIT (the last two were benchmarked by [1]). We note that each sample was assembled into a single contig with perfect or near-perfect quality.

Moreover, the results clearly show that one does not need any additional pre- and post-processing steps beyond quality trimming to obtain decent results, contrasting with the complicated assembly pipeline presented in [1] that included multiple steps of read binning, contamination filtering, and norovirus read filtering. All these steps might influence the final result and cause the degradation of assembly quality in general.

Since our approach only includes read trimming as a preprocessing step, and therefore can be directly compared to *pC* approach of [1], here we see that CORONASPADES was able to assemble seven out of eight samples into a contig longer than 7500 nt, and the contig length of the remaining sample is 7493 nt, which places this sample into a near-complete category. The original *pC* approach utilized METASPADES and MEGAHIT, which assembled four out of eight and five out of eight samples into contigs longer than 7500 nt, respectively. We summarized these results in Figure 1.

Further analysis might include checking the assembly graph to figure out if these assembly length differences are due to the assembler being over-aggressive in eliminating sequencing errors, or due, e.g., to uncollapsed viral quasi-species variation [16].
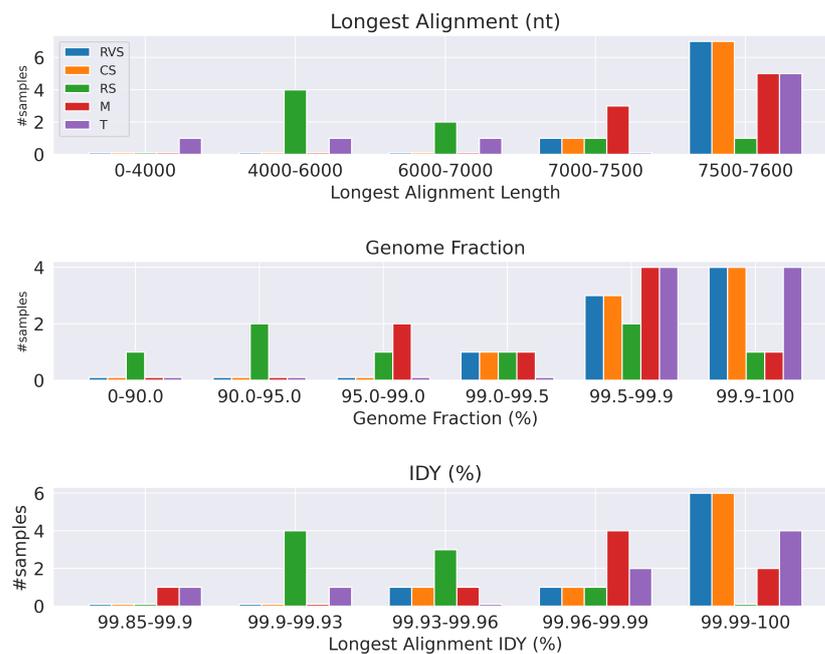
Figure 1. Summary of benchmarks for noroviral datasets. Tested tools are RNAVIRALSPADES (RVS), CORONASPADES (CS), RNASPADES (RS), MEGAHIT (M), and TRINITY (T). CORONASPADES and RNAVIRALSPADES have the best performance across all metrics.

Table 1. Benchmarking of assemblers RNAVIRALSPADES (RVS), CORONASPADES (CS), RNASPADES (RS), MEGAHIT (M), and TRINITY (T) on several noroviral datasets. The best results are outlined in bold.

		RVS	CS	RS	M	T
SRR8074276	Longest alignment (nt)	7538	7538	7538	**7548**	**7547**
	Genome fraction%	99.83	99.83	99.83	**99.96**	99.94
	Longest alignment IDY%	**99.95**	**99.95**	99.91	99.87	99.93
SRR9141472	Longest alignment (nt)	**7569**	**7569**	5282	7560	5848
	Genome fraction%	**100.0**	**100.0**	78.80	99.88	**100.0**
	Longest alignment IDY%	**100.0**	**100.0**	99.95	99.99	**100.0**
SRR9141473	Longest alignment (nt)	**7536**	**7536**	4,979	7487	6,899
	Genome fraction%	**100.0**	**100.0**	90.55	99.35	**100.0**
	Longest alignment IDY%	**100.0**	**100.0**	99.91	99.97	**100.0**
SRR9141474	Longest alignment (nt)	**7541**	**7541**	6,838	7516	**7542**
	Genome fraction%	**99.99**	**99.99**	99.99	99.65	**100.0**
	Longest alignment IDY%	**100.0**	**100.0**	99.91	99.99	**100.0**
SRR9141475	Longest alignment (nt)	7482	7493	7049	7440	**7534**
	Genome fraction%	99.28	99.43	99,08	98.72	**99.97**
	Longest alignment IDY%	**100.0**	**100.0**	99.96	**100.0**	**100.0**
SRR9141476	Longest alignment (nt)	**7533**	**7533**	5699	7526	**7533**
	Genome fraction%	**100.0**	**100.0**	**100.0**	99.90	**100.0**
	Longest alignment IDY%	**100.0**	**100.0**	99.98	**100.0**	99.98
SRR9141477	Longest alignment (nt)	**7554**	**7554**	5935	7467	3658
	Genome fraction%	**99.95**	**99.95**	91.20	98.79	99.84
	Longest alignment IDY%	**100.0**	**100.0**	99.93	99.96	99.97
SRR9141478	Longest alignment (nt)	**7540**	**7540**	6,592	**7540**	**7540**
	Genome fraction%	**100.0**	**100.0**	95.90	**100.0**	**100.0**
	Longest alignment IDY%	**99.99**	**99.99**	99.95	**99.99**	99.88

4. Discussion

The genome assembly task is a very challenging but quite well-studied computational problem. Multiple genome assemblers are available, and the choice of the particular tool is highly dependent on the input data properties and even the result desired. Nevertheless, even the proper choice of an assembler suitable for a given kind of input data cannot guarantee a complete genome assembly for complex datasets. However, we emphasize that even for datasets with low complexity, the researcher should choose an assembler carefully. A common problem that is seen in many papers is the usage of improper tools, improper options, or obsolete versions of the tools.

We showed that specialized RNA viral assemblers such as CORONASPADES and RNAVIRALSPADES were able to outperform metagenomic assemblers METASPADES and MEGAHIT and transcriptomic assemblers TRINITY and RNASPADES in the noroviral assembly task. While the two latter assemblers were designed to cope with the challenges of RNA sequencing data [13], their aim is completely different as they are trying to extract all possible valid transcript; therefore, their results might be suboptimal in the presence of multiple viral strains, quasispecies, and splice events [20].

We specifically outline that while CORONASPADES is able to improve viral assemblies, it requires a curated set of HMM profiles that matches a particular viral family. In addition to noroviral HMMs, we have already prepared coronaviral, HIV, and influenza HMMs (which are available for download from the CORONASPADES website), but most of the viral families were left uncovered. However, the procedure for creating a set of HMM profiles for a particular viral family is not a particularly complicated process.

Here, we provide some guidance on creating user-defined sets of HMMs. We recommend that one obtain a set of HMMs for the viral family of interest by extracting top-matching HMMs from a reliable source such as RVDB-prot-HMM [18]. However, due to the generic naming of the HMMs in the database, the hmmfetch from the HMMER3 suite [21] may not be the most straightforward option to use. Still, in many cases, it might be possible to query the provided annotation and create a subset of viral HMMs that might be used to supplement an assembly.

It might happen that the subset of HMMs extracted by annotation from RVDB-prot-HMM is too broad and can therefore match many extraneous sequences. In such a case, it would make sense to reduce the size of the set.

One way of doing this is to align HMMs from U-RVDB to one or several reference genomes from the viral family of interest using hmmscan. The top matching HMMs that have a significant match to the reference genome are usually relevant to the specific viral family. These HMMs can be extracted from U-RVDB using hmmfetch and used in CORONASPADES directly. Another option is to generate a custom set of HMMs if several reference genomes are available. The user can start by predicting coding sequences using tools such as Prodigal [22], GeneMark [23], or vGas [24], which are specifically designed for gene prediction and the extraction of coding sequences. Once the coding sequences have been extracted, they can be clustered using tools such as CD-HIT [25], UCLUST, MUSCLE [26], or MMSeqs [27], and HMMs can be built for each cluster using HMMER. This approach can be especially useful for studying highly diverse viral families, where standard HMMs may not be sufficient to accurately capture the diversity of the family.

Additionally, it is crucial to ensure that the resulting HMM set uniformly covers the reference genome, leaving no large gaps or uncovered segments. To achieve this, one can align the HMM set to the reference genome using hmmalign and check the alignment coordinates. It is worth noting that CORONASPADES can handle overlapping HMMs, although the obtained results are better interpreted when different HMMs do not overlap. A customized set of HMMs is a powerful tool that researchers can use to achieve more accurate and comprehensive viral genome assemblies, particularly when working with a large number of diverse datasets.

5. Conclusion

Our experiments showed that genome assembly using specialized tools and the latest versions of these tools yields better results in terms of both correctness and contiguity. Ad hoc assembly approaches end in worse results. Moreover, the labor costs associated with such approaches are much higher because suboptimal results force researchers to find a way to improve the initial results, which is usually harder than the assembly itself. Finally, this short note emphasizes the importance of specialized tool development and promotion.

References

1. Fuentes-Trillo, A.; Monzó, C.; Manzano, I.; Santiso-Bellón, C.; Andrade, J.d.S.R.d.; Gozalbo-Rovira, R.; García-García, A.B.; Rodríguez-Díaz, J.; Chaves, F.J. Benchmarking different approaches for Norovirus genome assembly in metagenome samples. *BMC Genom.* **2021**, *22*, 1–12. [CrossRef] [PubMed]
2. Edgar, R.C.; Taylor, J.; Lin, V.; Altman, T.; Barbera, P.; Meleshko, D.; Lohr, D.; Novakovsky, G.; Buchfink, B.; Al-Shayeb, B.; et al. Petabase-scale sequence alignment catalyses viral discovery. *Nature* **2022**, *602*, 142–147. [CrossRef]
3. Kawasaki, J.; Kojima, S.; Tomonaga, K.; Horie, M. Hidden viral sequences in public sequencing data and warning for future emerging diseases. *Mbio* **2021**, *12*, e01638-21. [CrossRef] [PubMed]
4. Sczyrba, A.; Hofmann, P.; Belmann, P.; Koslicki, D.; Janssen, S.; Dröge, J.; Gregor, I.; Majda, S.; Fiedler, J.; Dahms, E.; et al. Critical assessment of metagenome interpretation—A benchmark of metagenomics software. *Nat. Methods* **2017**, *14*, 1063–1071. [CrossRef] [PubMed]
5. Magoc, T.; Pabinger, S.; Canzar, S.; Liu, X.; Su, Q.; Puiu, D.; Tallon, L.J.; Salzberg, S.L. GAGE-B: An evaluation of genome assemblers for bacterial organisms. *Bioinformatics* **2013**, *29*, 1718–1725. [CrossRef] [PubMed]
6. Luo, W.; Friedman, M.S.; Shedden, K.; Hankenson, K.D.; Woolf, P.J. GAGE: Generally applicable gene set enrichment for pathway analysis. *BMC Bioinform.* **2009**, *10*, 1–17. [CrossRef]
7. Meyer, F.; Fritz, A.; Deng, Z.L.; Koslicki, D.; Lesker, T.R.; Gurevich, A.; Robertson, G.; Alser, M.; Antipov, D.; Beghini, F.; et al. Critical Assessment of Metagenome Interpretation: The second round of challenges. *Nat. Methods* **2022**, *19*, 429–440. [CrossRef]
8. Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P.A. metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **2017**, *27*, 824–834. [CrossRef]
9. Li, D.; Liu, C.M.; Luo, R.; Sadakane, K.; Lam, T.W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **2015**, *31*, 1674–1676. [CrossRef]
10. Roux, S.; Emerson, J.B.; Eloe-Fadrosh, E.A.; Sullivan, M.B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **2017**, *5*, e3817. [CrossRef]
11. Sutton, T.D.S.; Clooney, A.G.; Ryan, F.J.; Ross, R.P.; Hill, C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **2019**, *7*, 12. [CrossRef]
12. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef]
13. Bushmanova, E.; Antipov, D.; Lapidus, A.; Prjibelski, A.D. rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience* **2019**, *8*, giz100. [CrossRef]

14. Baaijens, J.A.; Aabidine, A.Z.E.; Rivals, E.; Schönhuth, A. De novo assembly of viral quasispecies using overlap graphs. *Genome Res.* **2017**, *27*, 835–848. [CrossRef]
15. Hunt, M.; Gall, A.; Ong, S.H.; Brener, J.; Ferns, B.; Goulder, P.; Nastouli, E.; Keane, J.A.; Kellam, P.; Otto, T.D. IVA: Accurate de novo assembly of RNA virus genomes. *Bioinformatics* **2015**, *31*, 2374–2376. [CrossRef]
16. Meleshko, D.; Hajirasouliha, I.; Korobeynikov, A. coronaSPAdes: From biosynthetic gene clusters to RNA viral assemblies. *Bioinformatics* **2022**, *38*, 1–8. [CrossRef]
17. Chan, M.C.; Kwan, H.S.; Chan, P.K. Structure and Genotypes of Noroviruses. In *The Norovirus*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 51–63. [CrossRef]
18. Bigot, T.; Temmam, S.; Pérot, P.; Eloit, M. RVDB-prot, a reference viral protein database and its HMM profiles [version 2; peer review: 2 approved]. *F1000Research* **2020**, *8*, 530. [CrossRef]
19. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef]
20. Viehweger, A.; Krautwurst, S.; Lamkiewicz, K.; Madhugiri, R.; Ziebuhr, J.; Hölzer, M.; Marz, M. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res.* **2019**, *29*, 1545–1554. [CrossRef]
21. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195. [CrossRef]
22. Hyatt, D.; Chen, G.L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **2010**, *11*, 1–11. [CrossRef] [PubMed]
23. Lukashin, A. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res.* **1998**, *26*, 1107–1115. [CrossRef] [PubMed]
24. Zhang, K.Y.; Gao, Y.Z.; Du, M.Z.; Liu, S.; Dong, C.; Guo, F.B. Vgas: A Viral Genome Annotation System. *Front. Microbiol.* **2019**, *10*, 184. [CrossRef] [PubMed]
25. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef]
26. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
27. Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028. [CrossRef]