1  **PALADIN:Protein Alignment for Functional Profiling Whole Metagenome Shotgun Data**

2

3  Anthony Westbrook[1], Jordan Ramsdell[1,2], Taruna Aggarwal[3], Louisa Normington[2], R. Daniel
4  Bergeron[1,2], W. Kelley Thomas[2,3], Matthew MacManes[2,3]

5

6  [1] University of New Hampshire, Department of Computer Science
7  [2] University of New Hampshire, Hubbard Center for Genome Studies
8  [3] University of New Hampshire, Department of Molecular Cellular and Biomedical Sciences

9

10  **Abstract**

11

12  Whole metagenome shotgun sequencing is a powerful approach for assaying the functional
13  potential of microbial communities. Presently, we lack tools that efficiently and accurately align
14  DNA reads against protein references, the technique necessary for constructing a functional
15  profile. Here, we present PALADIN—a novel modification of Burrows-Wheeler Aligner that
16  provides more accurate alignment and orders-of-magnitude improved efficiency by directly
17  mapping in protein space.

18

19  As high-throughput sequencing technologies improve, the analysis of microbial community
20  composition and function has rapidly advanced. Historically, this has mostly focused on
21  taxonomic surveys using a small number of phylogenetically informative genes such as the
22  small subunit of ribosomal RNA[1,2]. The ability to taxonomically profile communities provided new
23  insights into the role of microbiomes in human health[3,4], soil ecology[5,6] and environmental
24  remediation[7]. Nevertheless, the gene survey approach provides limited functional knowledge
25  because microorganisms with similar or even identical rRNA sequences often differ significantly
26  with respect to genomic content, and therefore may have vastly different functional roles in their
27  environment[8].

28  Functional profiling of microbial communities based on Whole Metagenome Shotgun (WMS)
29  sequencing data attempts to catalog the genes present in a community. An inventory of the
30  protein coding functions of a microbial community can be created by either matching the
31  individual reads to annotated reference databases or by assembling the reads and annotating
32  the resulting chromosomal segments[9]. Conventional methods such as BLAST are robust but
33  computationally intensive and techniques for rapidly mapping DNA reads to annotated reference
34  genes fail when the references within the curated databases diverge moderately from DNA
35  sequences of homologous genes in the metagenome sample. To mitigate these challenges,
36  researchers often turn to metagenome assembly and subsequent annotation which has
37  profound shortcomings, such as chimeric assembly of closely related sequences, strong bias
38  toward abundant organisms, and substantial human and computer resource requirements[2,9].
39  Therefore, current approaches are not sufficient to satisfy the requirements of researchers
40  attempting to understand functional metagenomics.

41

42  To improve the sensitivity with which functional profiling of metagenomics samples is
43  performed, we present PALADIN—an algorithm adapted from the popular BWA[11] mapping tool
44  (source code is available at https://github.com/twestbrookunh/paladin, see supplementary note
45  4 and figure 5 for implementation details). In brief, PALADIN identifies and translates six
46  possible open reading frames within each read, and maps these translated DNA sequence
47  reads to a protein reference allowing for rapid identification of functional metagenomic profiles.
48  By mapping in protein space, this method takes advantage of the general conservation of amino
49  acid sequences compared to the underlying DNA sequences.  Here we demonstrate the
50  application of this modified alignment algorithm for rapid and sensitive functional metagenomic
51  profiling using large scale WMS datasets. PALADIN reports mappings in standard SAM format,

52    and can generate a tab-delimited file from which additional information can be obtained,

53    including protein abundance and gene ontology.

54

55        To evaluate the performance of this novel protein space read mapper, we first generated

56    typical 250 basepair long paired-end reads using the standard Illumina error model for six well-

57    annotated bacterial genomes (*Pseudomonas fluorescens*, *Escherichia coli*, *Acidovorax avenae*,

58    *Micrococcus luteus*, *Halobacillus halophilus* and *Staphylococcus epidermidis)* using the read

59    simulation package ART[12]. The reads for the six genomes were pooled to create a mock-

60    metagenomic read dataset (see supplementary note 1). To establish a positive control, we used

61    PALADIN to map the combined reads to the protein sequences of the six original genomes and

62    to a filtered Swiss-Prot database (see supplementary note 2), which contains well-curated

63    protein sequences excluding the bacteria from which the reads were derived. To compare

64    PALADIN's relative mapping efficiency and accuracy to existing tools that align reads only in

65    DNA space, we mapped the nucleotide version of the reads to two reference types using BWA

66    and Novoalign (http://www.novocraft.com)—a DNA based read mapper with the capacity to map

67    reads to a degenerate nucleotide reference (see supplementary note 3). As expected, when

68    mapping reads to the genomes they were derived from, PALADIN and BWA map 98.29% and

69    96.02% of these reads respectively; By Contrast, Novoalign implementing the degenerate read

70    mapping methodology mapped 36.39% of the reads.  The poor performance of Novoalign

71    appears to stem from the fact that while it accepts degenerate bases in the references it does

72    not score them as positive matches during alignment.

73

74        By leveraging our prior knowledge about the genes corresponding to the simulated reads, we

75    evaluated the functional accuracy of mapping using three metrics—percentage of reads

76    mapped, Jaccard similarity coefficient[13], and the number of unique proteins found in Swiss-Prot.

77    To calculate the similarity coefficient, reads and their corresponding aligned targets were

78    assigned functionality using the standardized Gene Ontology (GO) language (see

79    supplementary note 5). Each GO term represents a vertex within a graph formation where

80    conditional edges join terms tracing back to one of three parent vertices: biological processes,

81    molecular function, and cellular component. For each read and its matching Swiss-Prot entry,

82    graphs were constructed by the GO term assignments directed back to their respective parent

83    vertices. The ratio of the intersection and the union of both graphs was used to determine the

84    Jaccard similarity coefficient as an alignment accuracy metric. The number of unique proteins

85    was determined based on each distinct Swiss-Prot ID the reads mapped to.

86

87     When we tested the ability of PALADIN to map mock-metagenomic reads to the Swiss-Prot

88    database, it mapped 30.65% of compared to 19.79% (BWA) and 0.56% (Novoalign), while all

89    three systems mapped with extremely high accuracy (correct functional assignment) measured

90    by the similarity index (Table 1). These results suggest that mapping in protein space as

91    implemented in PALADIN is even more accurate than existing solutions while using a larger

92    proportion of the available shotgun metagenomics dataset.

93

94

95

96

97

98

99

|  |  | Original Genomes (CDS) | | | Filtered Swiss-Prot | | | Filtered UniRef90 | |
|---|---|---|---|---|---|---|---|---|---|
|  |  | **BWA** | **Novo** | **PALADIN** | **BWA** | **Novo** | **PALADIN** | **BWA** | **PALADIN** |
| Test | Reads mapped % | 96.02 | 36.39 | 98.29 | 19.79 | 0.56 | 30.65 |  |  |
|  | Similarity index | 0.86 | 0.92 | 0.93 | 0.81 | 0.81 | 0.83 |  |  |
|  | Unique proteins | 20448 | 19932 | 22127 | 16808 | 1215 | 39496 |  |  |
| Lung | Reads mapped % |  |  |  |  |  |  | 73.9 | 86.85 |
| Gut | Reads mapped % |  |  |  |  |  |  | 74.6 | 79.9 |
| Soil | Reads mapped % |  |  |  |  |  |  | 66.8 | 71.25 |

Table 1. Mapping efficiency of PALADIN, BWA, and Novoalign against multiple read sets. Positive control was established by aligning the simulated reads against the original six test genomes. This simulated read set was then aligned against the filtered Swiss-Prot reference, resulting in PALADIN performing best in percentage of reads mapped, similarity index, and number of aligned proteins. Additionally, three empirically derived read sets (lung (BioProject:PRJNA71831), gut (BioSample:SAMN00037421), and soil (MG-RAST:4520320.3)) were aligned against the filtered UniRef90 reference, each resulting in PALADIN scoring better than both BWA and Novoalign.

100

101 While mapping mock reads to the well annotated SwissProt database allows us to assay
102 accuracy via the use of GO-term similarity, it is a relatively small dataset with limited
103 representation of both functional and taxonomic breadth. A more ideal reference would be the
104 UniRef90 database which contains taxonomically diverse sequences clustered at 90%
105 sequence identity. In a process identical to above, we mapped translated DNA reads from three
106 published WMS dataset from different environments to the Uniref90 proteins using PALADIN,
107 and untranslated reads to the corresponding nucleotide sequences from Uniref90 proteins using
108 BWA. With mapping accuracy established in the first set of experiments, the enhanced mapping
109 rate using PALADIN versus BWA (Table 1) translates to increased resolution of functional
110 profiling.

111

112 To contrast the differences in computational efficiency between PALADIN and conventional
113 protein alignment tools, we mapped the reads of a dataset consisting of nearly 240,000,000
114 reads against the Uniref90 database with PALADIN using 28 cores on a high-end workstation.
115 The computation finished after 31 hours for an efficiency of approximately 128,000
116 reads/minute. We then extracted 8,000 of these sequences from the dataset and performed an
117 alignment with BLAST[10] using the same hardware environment and resource availability.
118 BLAST completed execution in 8.5 hours with an efficiency of about 16 reads/minute. Given the
119 linear time complexity associated with BLAST, we estimate that an execution run against the full
120 dataset would take about 29 years, approximately 8,000 times longer than PALADIN.

121

122 In summary, we present PALADIN, a tool for accurate functional characterization of
123 metagenomic samples that is orders of magnitude faster than existing approaches. This
124 significant improvement in efficiency affords researchers unprecedented opportunity to gain
125 detailed and novel insight into microbial communities. Additionally, by constructing this
126 approach upon a widely used alignment algorithm, reliability and usability are inherently
127 increased, which promotes faster adoption and easier incorporation into existing pipelines.
128 Finally, the reduction in required computational resources creates a more cost-effective solution,
129 thereby increasing viability of analysis capabilities in environments where economic pressures
130 are present. Given these aspects, PALADIN may potentially aid in any number of evolving
131 fields that depend on functional characterization, including personalized medicine, biodefense,
132 environmental remediation, transcriptomics, and the study of emerging pathogens.

133

134 **Acknowledgements**

135 **References**

136 1. Tap, J. *et al*. *Environ. Microbiol*. **11**, 2574-2584 (2009).

137 2. Scholz, M.B. *et al*. *Curr Opin Biotechnol*. **23**, 9-15 (2012).

138 3. Qin, J. *et al*. *Nature*. **464**, 59-67 (2010).

139 4. Cho, I., Blaser, M.J. *Nat. Rev. Genet*. **13**, 260–270 (2012).

140 5. Rinke, C. *et al*. *Nature*. **499**, 431-437 (2013).

141 6. Hultman, J. *et al*. *Nature*. **521**, 208-212 (2015).

142 7. Fierer, N. *et al*. *Science*. **342**, 621-624 (2013).

143 8. Sentausa, E., Fournier P.E. *Clin Microbiol & Infect*. **19**, 790-795 (2013).

144 9. Nagarajan, N., Pop, M. *Nat. Rev. Genet*. **14**, 157–167 (2013).

145 10. Altschul, S.F. *et al. J. Mol. Bio.* **215**, 403-410 (1990).
146
147 11. Li, H., Durbin, R. *Bioinformatics*. **25**, 1754-1760 (2009).
148
149 12. Huang, W., Li, L., Myers, J.R., Marth, G.T. *Bioinformatics*. **28**, 593-594 (2012).
150
151 13. Jaccard, P. *Bulletin de la Société Vaudoise des Sciences Naturelles 37: 547–579* (1901)

152