Supplementary Materials to "Orphelia: predicting genes in metagenomic sequencing reads"

Katharina J. Hoff [1]*, Thomas Lingner [1,2], Peter Meinicke [1] and Maike Tech [1]

April 8, 2009

[1]Abteilung Bioinformatik, Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany, E-Mails: {katharina, peter, maike}@gobics.de [2]Center for Genomic Regulation, Comparative Bioinformatics Research Group, Biomedical Research Park de Barcelona, c/Dr. Aiguader 88, 08003 Barcelona, Spain, E-Mail: thomas.lingner@crg.es

Methods

Parameters of sequencing read simulation with MetaSim

The sequence database used by MetaSim was downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.fna.tar.gz on 02/12/09. For all simulations, the following taxon-profile was used:

```
1    taxid    243232 # Methanocaldococcus_jannaschii_DSM_2661
1    taxid    107806 # Buchnera aphidicola str. APS (Acyrthosiphon pisum)
1    taxid    224325 # Archaeoglobus fulgidus DSM 4304
1    taxid    348780 # Natronomonas pharaonis DSM 2160
1    taxid    74546 # Prochlorococcus marinus str. MIT 9312
1    taxid    292805 # Wolbachia endosymbiont strain TRS of Brugia malayi
1    taxid    85963 # Helicobacter pylori J99
1    taxid    224308 # Bacillus subtilis subsp. subtilis str. 168
```

*to whom correspondence should be addressed

```
1    taxid    511145 # Escherichia coli str. K-12 substr. MG1655
1    taxid    194439 # Chlorobium tepidum TLS
1    taxid    306537 # Corynebacterium jeikeium K411
1    taxid    272560 # Burkholderia pseudomallei K96243 chromosome 1
1    taxid    272560 # Burkholderia pseudomallei K96243 chromosome 2
```

Sanger reads

Sanger reads were sampled with the following general simulation parameters:

```
Number Of Reads / Mate Pairs=45142
Error Model=Sanger

Sanger Error Model Read Length Parameters=
Distribution: Normal
Mean: 700.0
2nd parameter: 100.0

Sanger Error Model Configuration=
Mate Pair Probability: 1.0
Proportion of Deletion Errors: 0.2
Proportion of Insertion Errors: 0.2
Proportion of Substitution Errors: 0.6000000000000001

Sanger Error Model DNA Clone Parameters=
Distribution: Normal
Mean: 5000.0
2nd parameter: 500.0
```

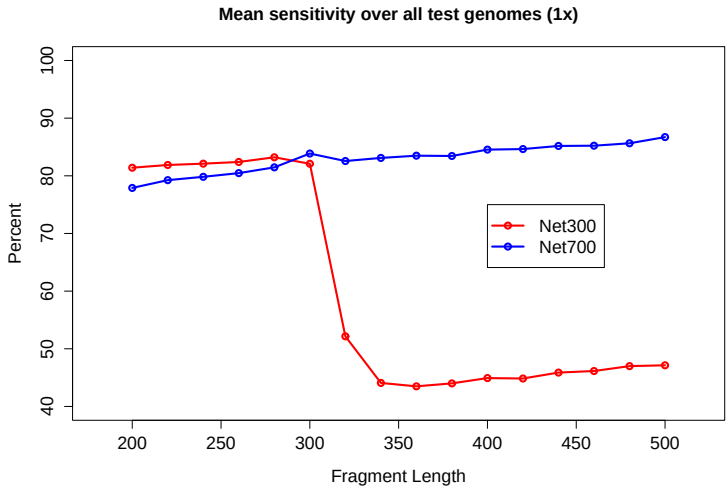
```
Combine All Files=false
Uniform Sequence Weights=false
Number Of Threads=1
Write FastA=true
Compress Output Files=false
```

For five different simulation scenarios, different error rates at the beginning (beg.) and end of read were used:

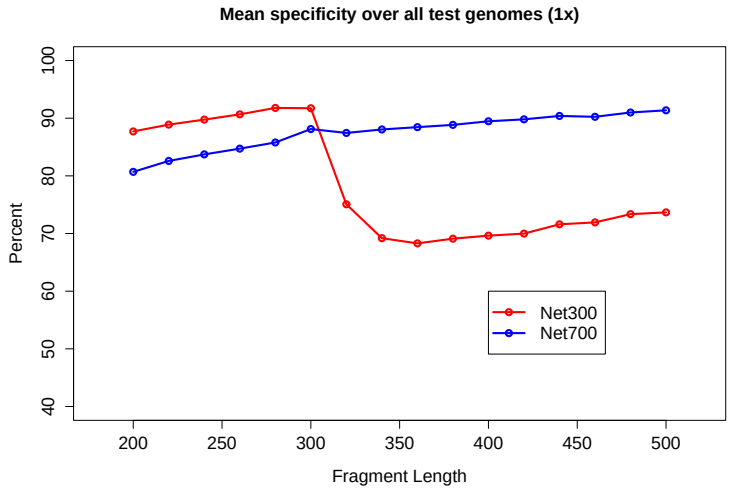
Scenario No.	error rate beg.	error rate end
I	0	0
II	0.01	0.02
III	0.001	0.002
IV	0.0001	0.0002
V	0.00001	0.00002

Supplementary Figures

Supplementary Figure 1: Sensitivity of Orphelia with Net300 and Net700 on fragments ranging from 200 to 500 bp length (sampled to a 1-fold genome coverage).



Supplementary Figure 2: Specificity of Orphelia with Net300 and Net700 on fragments ranging from 200 to 500 bp length (sampled to a 1-fold genome coverage).



Supplementary Tables

Supplementary Table 1: Gene prediction performance (on amino acid level) of Orphelia Net700 on simulated Sanger reads with different error rates.

Error rate beginning of read	Error rate end of read	Sensitivity	Specificity
0	0	95.3	96.3
1×10^{-5}	2×10^{-5}	95.3	96.3
1×10^{-4}	2×10^{-4}	94.5	95.9
1×10^{-3}	2×10^{-3}	87.6	93.3
1×10^{-2}	2×10^{-2}	47.2	75.3