NanoMGT

Marker gene typing of low complexity

mono-species metagenomic samples using noisy long

reads

Metagenomics

- 1.) What is in the sample
- 2.) At what abundances
- 3.) Which organisms did genes originate from

Long reads: Solve A LOT of metagenomic-related challenges.

Metagenomic binning / Taxonomic classification approaches

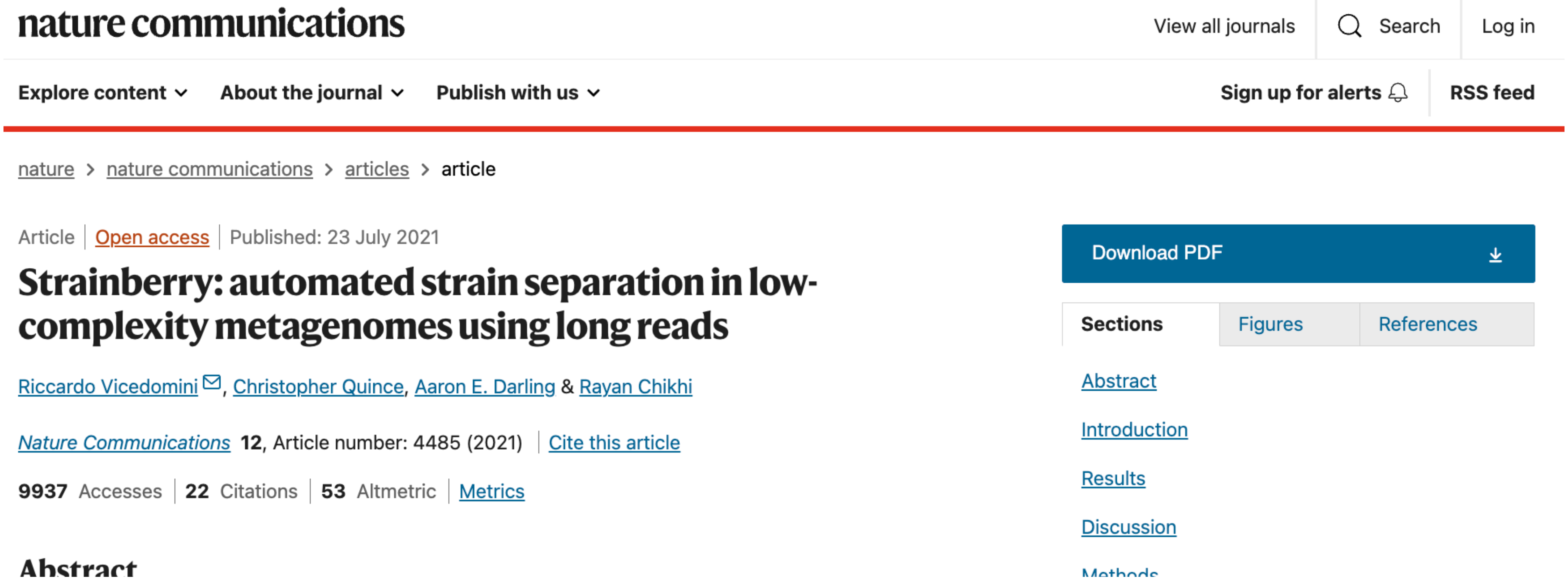
ASSEMBLY-BASED

ALIGNMENT-BASED

LCA, MASH
SKETCHES, K-MERS
ETC.

The ideal metagenomic pipeline

- 1.) Assembly free species-level binning
- 2.) NanoMGT run on each species-level bin
- 3.) Phasing of strains using NanoMGT results



nature communications

Article | Open access | Published: 23 July 2021

Strainberry: automated strain separation in low-complexity metagenomes using long reads

Riccardo Vicedomini, Christopher Quince, Aaron E. Darling & Rayan Chikhi

Nature Communications **12**, Article number: 4485 (2021) | Cite this article

9937 Accesses | **22** Citations | **53** Altmetric | Metrics

Abstract

What we want

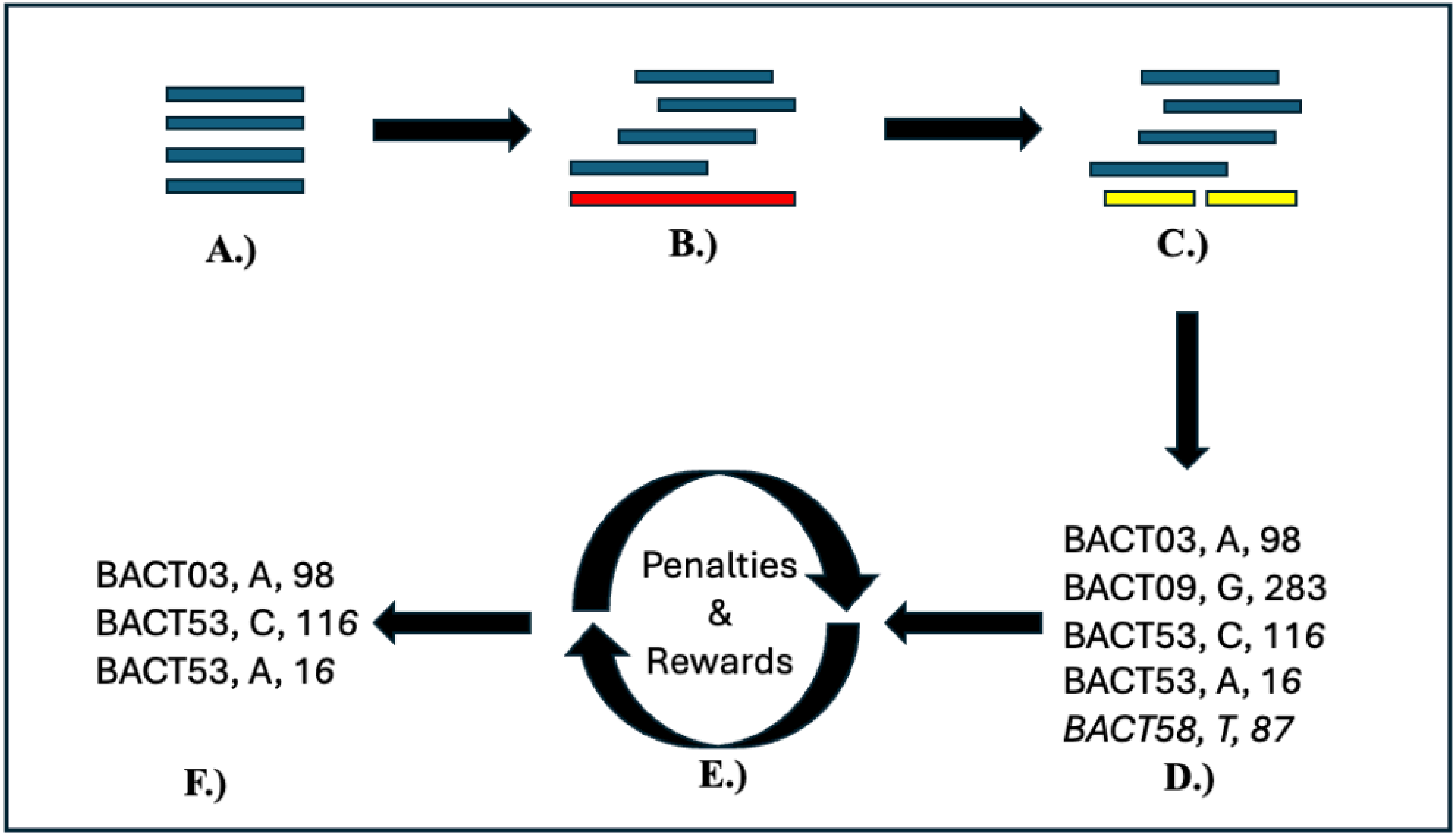
1.) Isolate reads for a species

2.) Type variant positions within the bin (NanoMGT)

3.) Determine is more than one strain is likely to be present

Existing variant callers for long read metagenomic data

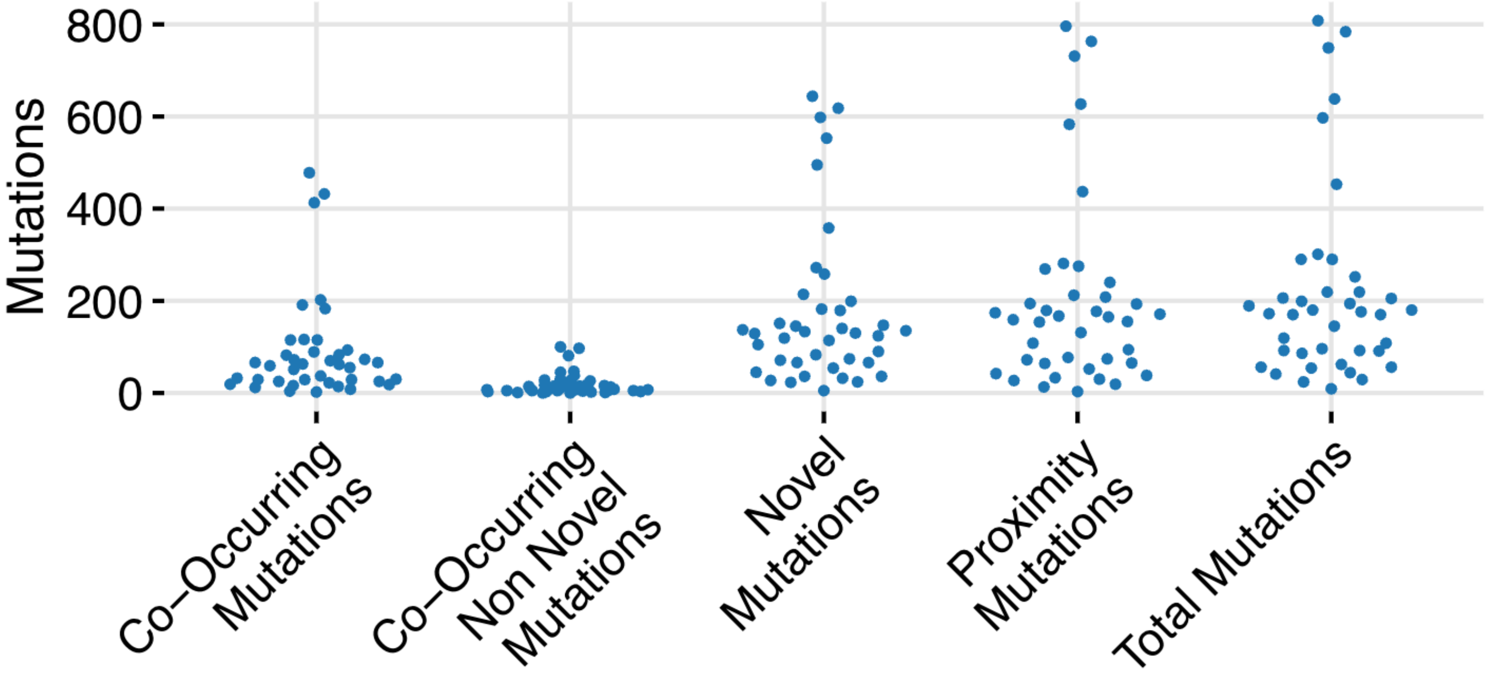
Medaka: Refines whole genomes, can't solve the problem	Trained neural network
LongShot: Diploid variant caller, can identify some variants with high depth, but generally doesn't ID much.	Preset thresholds, density filter
ConFindr: rMLST-based minority variant caller	Proximity trimming, preset thresholds



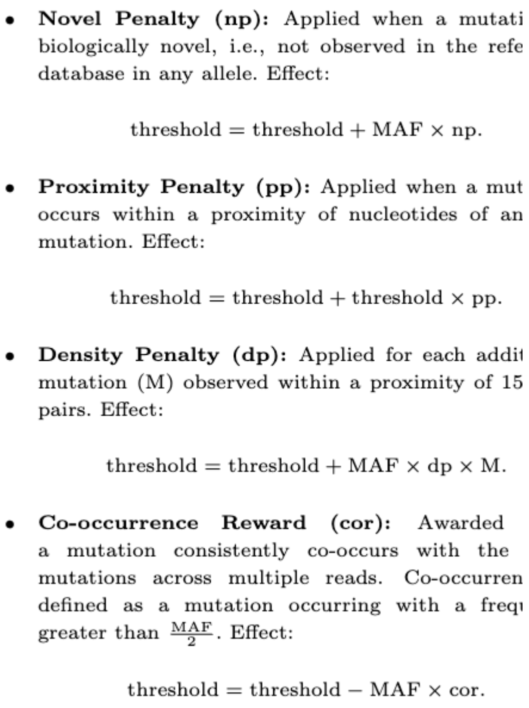
BACT03, A, 98
BACT09, G, 283
BACT53, C, 116
BACT53, A, 16
BACT58, T, 87

BACT03, A, 98
BACT53, C, 116
BACT53, A, 16

ONT error profiles



NanoMGT Algorithm



- **Novel Penalty (np):** Applied when a mutati biologically novel, i.e., not observed in the refe database in any allele. Effect:

$$\text{threshold} = \text{threshold} + \text{MAF} \times \text{np}.$$

- **Proximity Penalty (pp):** Applied when a mut occurs within a proximity of nucleotides of an mutation. Effect:

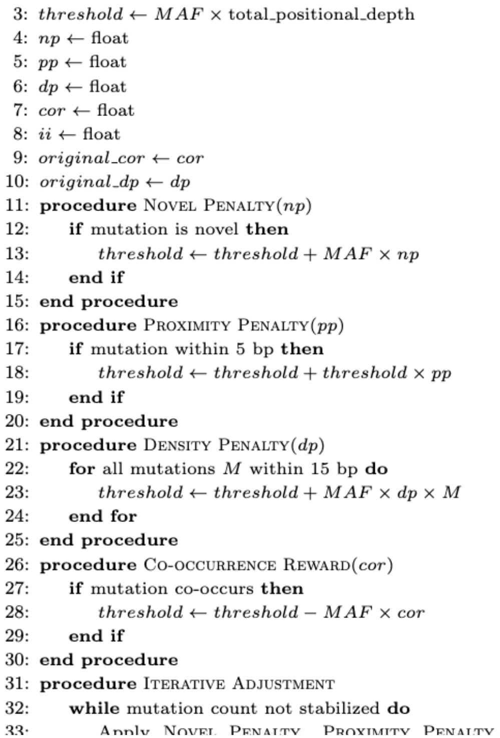
$$\text{threshold} = \text{threshold} + \text{threshold} \times \text{pp}.$$

- **Density Penalty (dp):** Applied for each addi mutation (M) observed within a proximity of 15 pairs. Effect:

$$\text{threshold} = \text{threshold} + \text{MAF} \times \text{dp} \times \text{M}.$$

- **Co-occurrence Reward (cor):** Awarded a mutation consistently co-occurs with the mutations across multiple reads. Co-occurren defined as a mutation occurring with a freq greater than $\frac{\text{MAF}}{2}$. Effect:

$$\text{threshold} = \text{threshold} - \text{MAF} \times \text{cor}.$$



```
$threshold \leftarrow MAF \times$ total_positional_depth
$np \leftarrow$ float
$pp \leftarrow$ float
$dp \leftarrow$ float
$cor \leftarrow$ float
$ii \leftarrow$ float
$original\_cor \leftarrow cor$
$original\_dp \leftarrow dp$
procedure NOVEL PENALTY($np$)
    if mutation is novel then
        $threshold \leftarrow threshold + MAF \times np$
    end if
end procedure
procedure PROXIMITY PENALTY($pp$)
    if mutation within 5 bp then
        $threshold \leftarrow threshold + threshold \times pp$
    end if
end procedure
procedure DENSITY PENALTY($dp$)
    for all mutations $M$ within 15 bp do
        $threshold \leftarrow threshold + MAF \times dp \times M$
    end for
end procedure
procedure CO-OCCURRENCE REWARD($cor$)
    if mutation co-occurs then
        $threshold \leftarrow threshold - MAF \times cor$
    end if
end procedure
procedure ITERATIVE ADJUSTMENT
    while mutation count not stabilized do
        Apply NOVEL PENALTY, PROXIMITY PENALTY
```

Data

Combined dataset: 39 isolates, six species.

Clean dataset: 24 isolates

Contaminated dataset: 15 isolates

Parameter search

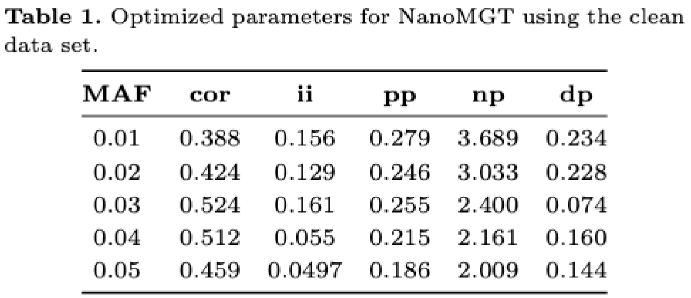


Table 1. Optimized parameters for NanoMGT using the clean data set.

MAF	cor	ii	pp	np	dp
0.01	0.388	0.156	0.279	3.689	0.234
0.02	0.424	0.129	0.246	3.033	0.228
0.03	0.524	0.161	0.255	2.400	0.074
0.04	0.512	0.055	0.215	2.161	0.160
0.05	0.459	0.0497	0.186	2.009	0.144

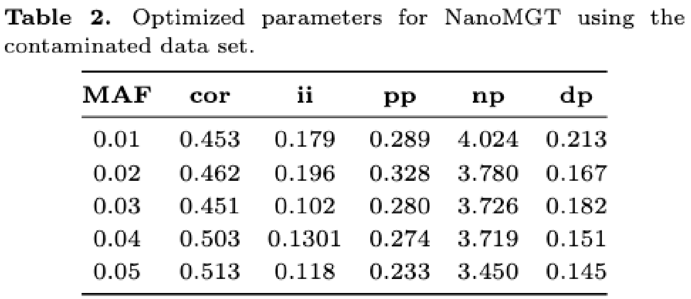


Table 2. Optimized parameters for NanoMGT using the contaminated data set.

MAF	cor	ii	pp	np	dp
0.01	0.453	0.179	0.289	4.024	0.213
0.02	0.462	0.196	0.328	3.780	0.167
0.03	0.451	0.102	0.280	3.726	0.182
0.04	0.503	0.1301	0.274	3.719	0.151
0.05	0.513	0.118	0.233	3.450	0.145

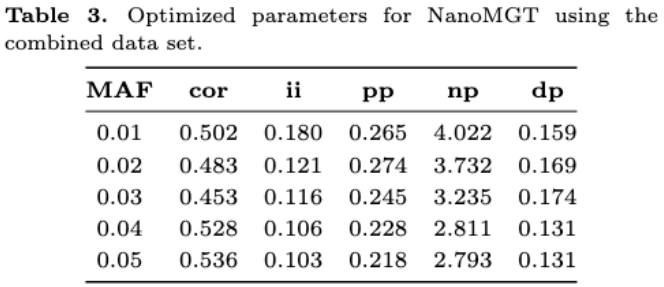
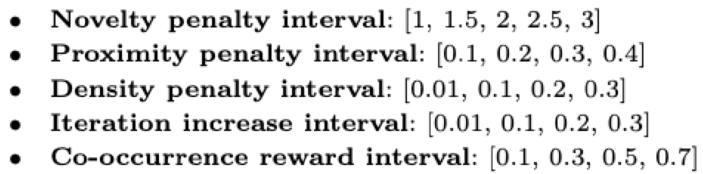


Table 3. Optimized parameters for NanoMGT using the combined data set.

MAF	cor	ii	pp	np	dp
0.01	0.502	0.180	0.265	4.022	0.159
0.02	0.483	0.121	0.274	3.732	0.169
0.03	0.453	0.116	0.245	3.235	0.174
0.04	0.528	0.106	0.228	2.811	0.131
0.05	0.536	0.103	0.218	2.793	0.131



- **Novelty penalty interval**: [1, 1.5, 2, 2.5, 3]
- **Proximity penalty interval**: [0.1, 0.2, 0.3, 0.4]
- **Density penalty interval**: [0.01, 0.1, 0.2, 0.3]
- **Iteration increase interval**: [0.01, 0.1, 0.2, 0.3]
- **Co-occurrence reward interval**: [0.1, 0.3, 0.5, 0.7]

Variant positions in the data

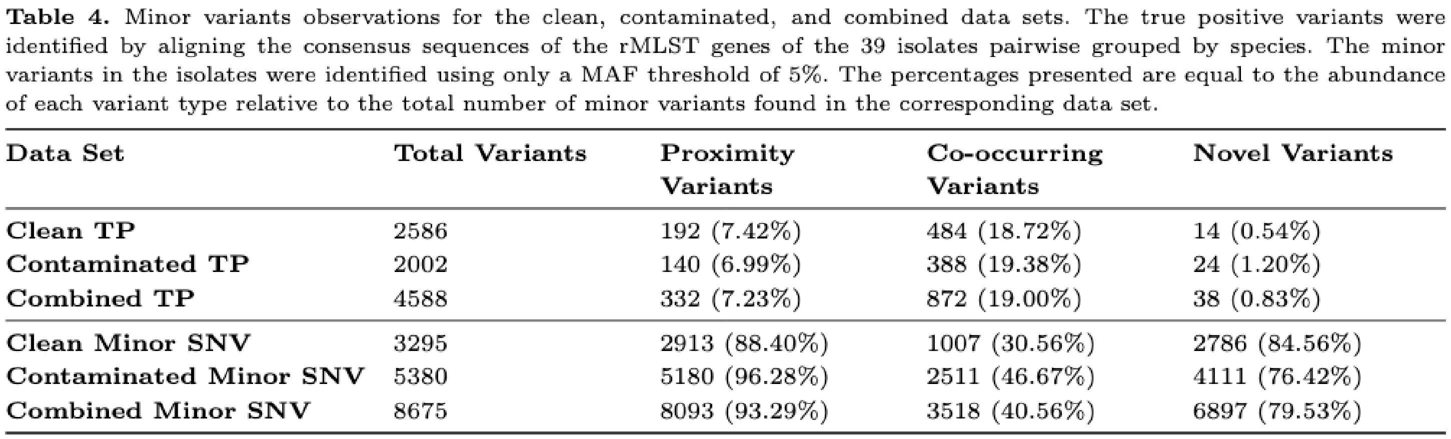
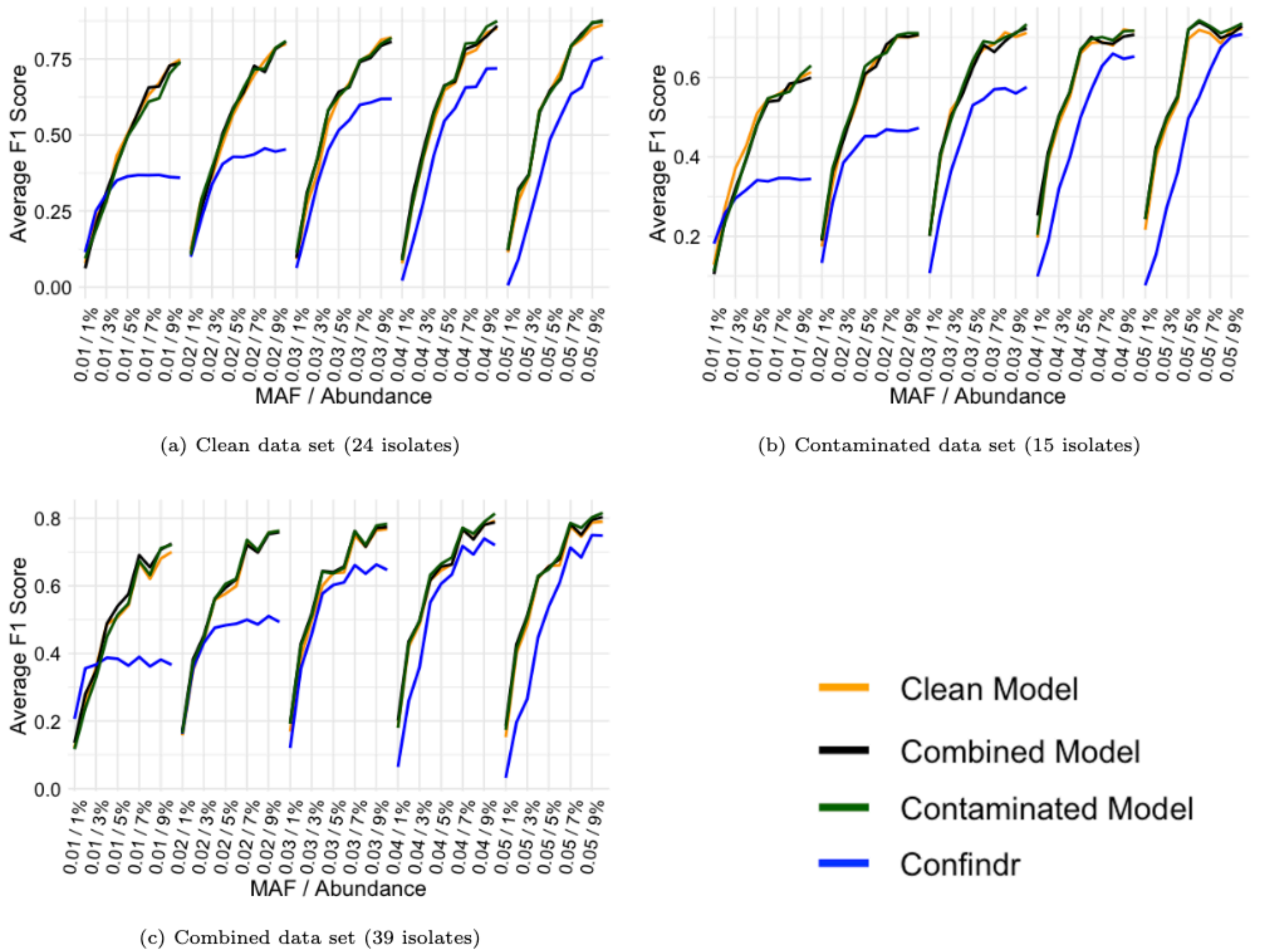


Table 4. Minor variants observations for the clean, contaminated, and combined data sets. The true positive variants were identified by aligning the consensus sequences of the rMLST genes of the 39 isolates pairwise grouped by species. The minor variants in the isolates were identified using only a MAF threshold of 5%. The percentages presented are equal to the abundance of each variant type relative to the total number of minor variants found in the corresponding data set.

Data Set	Total Variants	Proximity Variants	Co-occurring Variants	Novel Variants
Clean TP	2586	192 (7.42%)	484 (18.72%)	14 (0.54%)
Contaminated TP	2002	140 (6.99%)	388 (19.38%)	24 (1.20%)
Combined TP	4588	332 (7.23%)	872 (19.00%)	38 (0.83%)
Clean Minor SNV	3295	2913 (88.40%)	1007 (30.56%)	2786 (84.56%)
Contaminated Minor SNV	5380	5180 (96.28%)	2511 (46.67%)	4111 (76.42%)
Combined Minor SNV	8675	8093 (93.29%)	3518 (40.56%)	6897 (79.53%)

Performance results



(a) Clean data set (24 isolates)

(b) Contaminated data set (15 isolates)

(c) Combined data set (39 isolates)

Fig. 3. Average F1 performance across the simulated multistrain samples from different data sets using Confindr and NanoMGT run with all 3 parameter models. The F1-score was calculated for MAF values running from 0.01 to 0.05 (presented as whole percentage integers in the plot) in combination with the abundance of the minority isolates in the multistrain samples running from 1%-10%. Only every other data point on the x-axis is displayed to enhance readability.

Conclusions

- Threshold-based approach: Works much better than proximity filtering.
- As MAJOR indicator of errors is biological novelty.
- In the future, LLMs capable of understanding nucleotide language could prove very powerful as error correctors.