Krisp: A python package for designing CRISPR-based and amplification-based diagnostic assays using whole genome data

Zachary S.L. Foster, Andrew S. Tupper, Caroline M. Press, Jeff H. Chang, and Niklaus J. Grünwald

Summary:

CRISPR-Cas combined with isothermal amplification is used for diagnostic assays for sequence-specific detection of DNA/RNA. These assays have similar sensitivity to qPCR and can be made into inexpensive test strips. Finding diagnostic loci flanked by conserved regions where primers can be designed requires extensive analyses of genome sequences. We developed the python package krisp to find primers and diagnostic sequences that differentiate groups of samples from each other using either unaligned genome sequences or a variant call format (VCF) file as input. Krisp can process large data sets by using efficient algorithms that run in near linear time and in parallel. Results have been validated in the laboratory with the design of a SHERLOCK assay to detect the sudden oak death pathogen Phytophthora ramorum. Krisp is released open source at https://github.com/grunwaldlab/krisp with the documentation needed to quickly design diagnostic assays.

Methods:

Krisp has two principal functions: krisp_fasta and krisp_vcf. Both functions:

- Have a command line interface
- Run in parallel, use minimal RAM, and have execution times that scale linearly with the size of input data (Figure 1)
- Use Primer3 to filter results based on the presence of potential primer sites (Figure 2)
- Return CSV output for downstream analysis and human-readable alignments for manual inspection (Figures 2 and 3)

Krisp_fasta infers diagnostic sites from whole genome assemblies by searching for k-mers that distinguish groups and have conserved sequence where primers can be designed.

Krisp_vcf analyses a sliding window of variants in a VCF file. When one or more diagnostic variants are found surrounded by conserved sequence, the sequence for each group is inferred by applying variants to the reference sequence.

Validation:

Krisp_fasta and krisp_vcf were validated in the lab by using each to design a SHERLOCK diagnostic assay to differentiate P. ramorum from closely related Phytophthora species. For krisp_fasta, whole genome sequences from 5 P. ramorum isolates were compared to 6 other Phytophthora species. For krisp_vcf, a VCF file containing variants of 589 P. ramorum samples and 7 samples of closely related species was used. A promising diagnostic site was selected that occurred in the output of both methods and tested in the lab. Our results suggest the assay works as expected. All P. ramorum samples representing the 4 major lineages resulted in fluorescence whereas negative controls, including 6 other Phytophthora species, did not. (Figure 4).

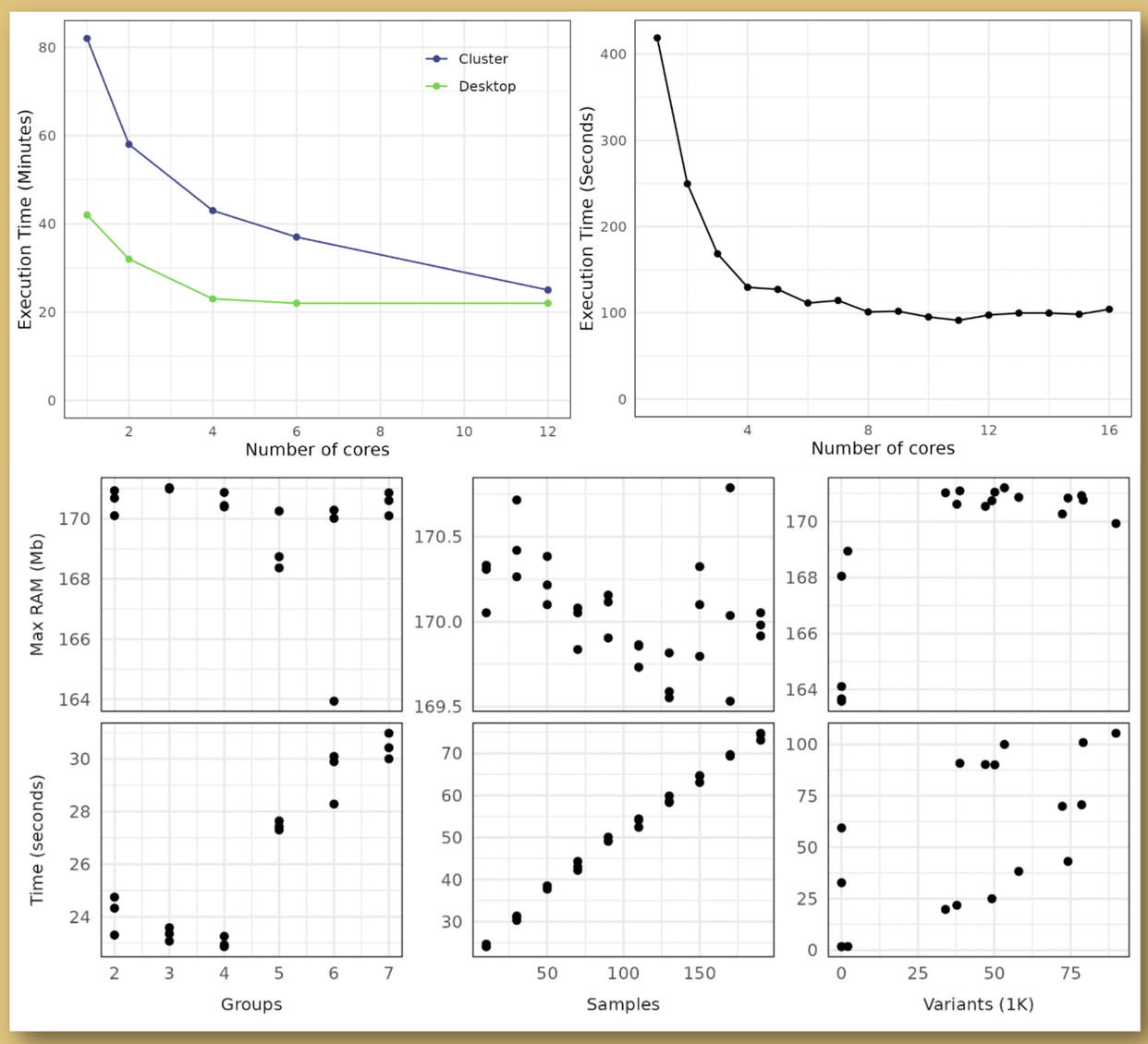

Figure 1: The effects of number of cores, variant count, sample count, and number of groups being distinguished on the execution time and maximum RAM usage of krisp_vcf.

```
## Phyram_PR-102_s0001:209731-209856 is diagnostic for NA2

Reference : gtcccgtcaccgtatatatgtactaaacgca...gtgggtagcatactgacgacgagaagt  C  agt...ctggttaggacagttaaatgtaccGagag
      EU1 : ..............................................................  C9 ..............................R....
      NA1 : ..............................................................  C6 ..................................
      NA2 : ..............................................................<T11>..................................
   oligos : gtcccgtcaccgtatatatgtactaaacg  ...  gggtagcatactgacgacgagaagt  T  ag ...  ggttaggacagttaaatgtaccgagag
            └──────── Left primer ────────┘  ...  └──────────── crRNA ─────────────┘ ...  └────── Right primer ──────┘

Primer statistics:
 Direction  Penalty  Sequence                          Tm        Gc Percent  Self Any Th  Self End Th  Hairpin Th ...
 Forward    3.42431  gtcccgtcaccgtatatatgtactaaacg  62.92431  44.82759    7.87208      10.60241     0.0        ...
 Reverse    3.60529  ctctcggtacatttaactgtcctaacc    61.10529  44.44444    1.18023      3.98189      0.0        ...

Pair statistics:
 Penalty  Compl Any Th  Compl End Th  Product Size  Product Tm  Product Tm Oligo Tm Diff  T Opt A            ...
 7.0296   2.27423       0.0           126           83.22136    22.11607                  61.68654           ...
```

Figure 2: Example of the human-readable output of krisp with Primer3 statistics. Diagnostic sites are indicated with angle brakets (e.g., <T11>) and the number of samples supporting each allele are shown. Variable sites are shown as IUPAC ambiguity codes if they are in non-diagnostic sites.

	A	B	C	D	E	F	G	H	I	J	K	
1	group	chrom	n_diag	reg_from	reg_to	diag_from	diag_to	fwd_from	fwd_to	rev_from	rev_to	seq_adj_left
2	EU1	Phyram_PR-102_s0001	1	179266	179556	179391	179418	179344	179372	179442	179466	acacgccgtactgcgccaacactttaccggaatcact
3	NA1	Phyram_PR-102_s0001	1	182656	182956	182781	182808	182694	182723	182816	182842	ggcggcttggaccgagtatgaacctccattcaatactc
4	EU1	Phyram_PR-102_s0001	2	182656	182958	182783	182810	182694	182723	182816	182842	ggcgacttggaccgagtatgaacctccattcaatactc
5	EU1	Phyram_PR-102_s0001	2	182656	182959	182784	182811	182694	182723	182816	182842	ggcgacttggaccgagtatgaacctccattcaatactc

Figure 3: Example of the first few columns of CSV output of krisp**.**

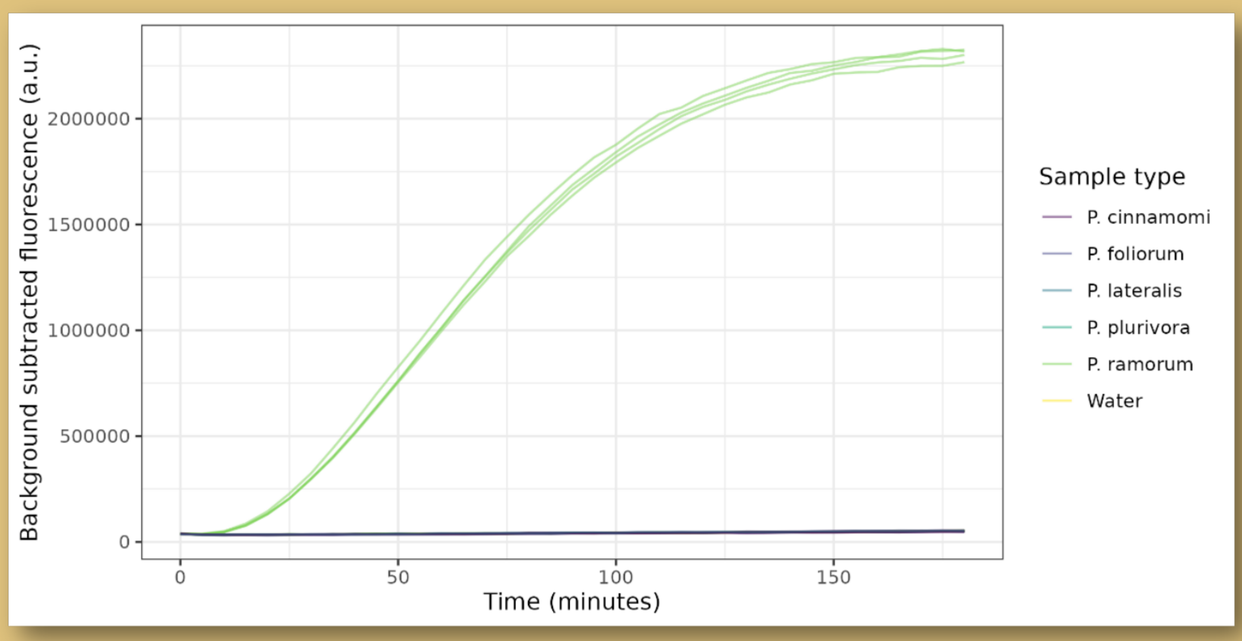

Figure 4: Lab validation of a proof-of-concept SHERLOCK assay predicted by both krisp_vcf and krisp_fasta to identify P. ramorum.

This work was supported by grants from USDA ARS (2072-22000-041-000-D), USDA NIFA (2023-67013-39918) and USDA APHIS to NG.