



Next Generation Sequencing as a New Universal Novel Pathogen Discovery Tool



Conceptual Outline



Detection of unknown pathogens

Sample



1 day?



NG sequencing





Detection of unknown pathogens

Sample



1 day?



NG sequencing



3-5 days



Set of reads

GATCTCATCTAGCATGAAGT
AAATCTCATCTAGCATGAAG
CATCTCAAATCTAGCATGAA
TGATCTCATCTAGCATGAAG
CCTCTCATCTAGCATGAAGT
AGATCTCAAATCTAGCATGA
GATCTCATCTAGCATGAAGT
TTTCTCATCTAGCATGAAGT
GATCTCATCTAGCATGAAGT
AAATCTCATCTAGCATGAAG
CATCTCAAATCTAGCATGAA
TGATCTCATCTAGCATGAAG
CCTCTCATCTAGCATGAAGT
AGATCTCAAATCTAGCATGA
GATCTCATCTAGCATGAAGT
TTTCTCATCTAGCATGAAGT



Detection of unknown pathogens

Sample



1 day?



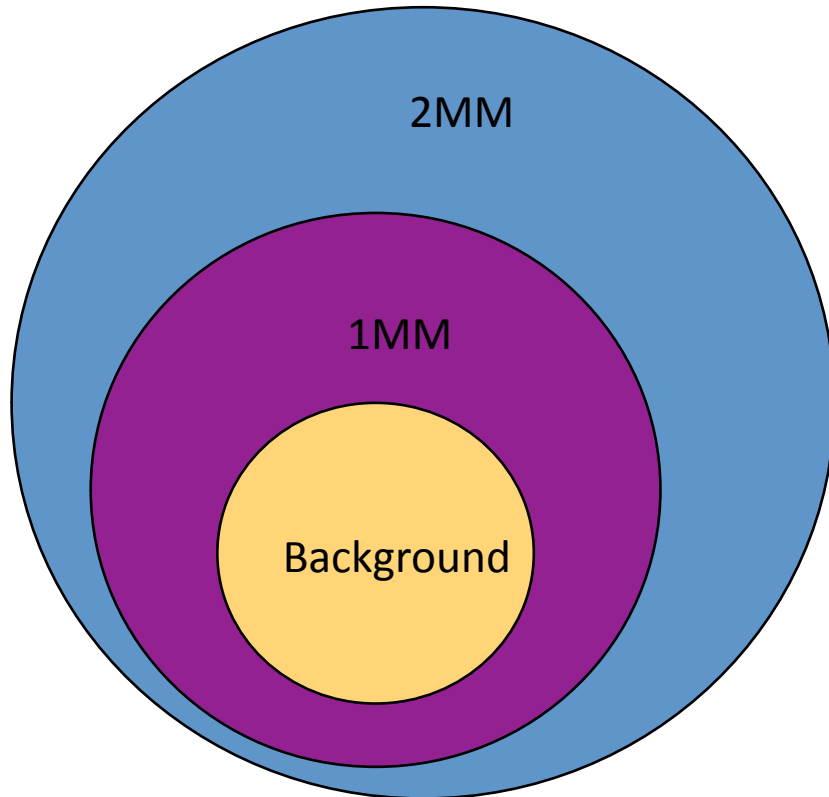
NG sequencing



3-5 days



Set of reads



GATCTCATCTAGCATGAAGT
AAATCTCATCTAGCATGAAG
CATCTCAAATCTAGCATGAA
TGATCTCATCTAGCATGAAG
CCTCTCATCTAGCATGAAGT
AGATCTCAAATCTAGCATGA
GATCTCATCTAGCATGAAGT
TTTCTCATCTAGCATGAAGT
GATCTCATCTAGCATGAAGT
AAATCTCATCTAGCATGAAG
CATCTCAAATCTAGCATGAA
TGATCTCATCTAGCATGAAG
CCTCTCATCTAGCATGAAGT
AGATCTCAAATCTAGCATGA
GATCTCATCTAGCATGAAGT
TTTCTCATCTAGCATGAAGT



Detection of unknown pathogens

Sample



1 day?



NG sequencing

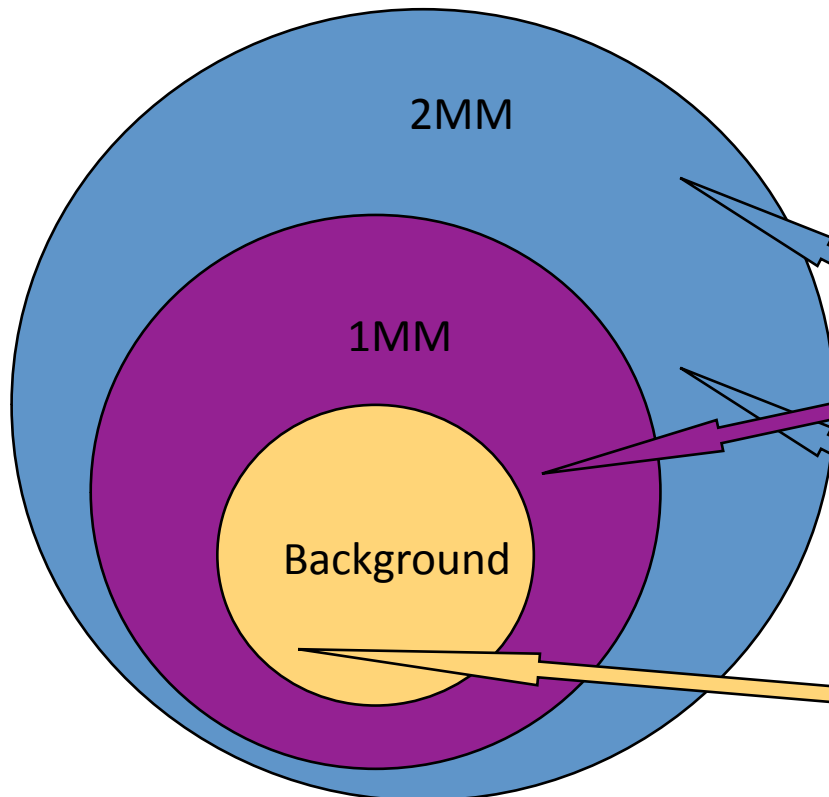


3-5 days



Set of reads

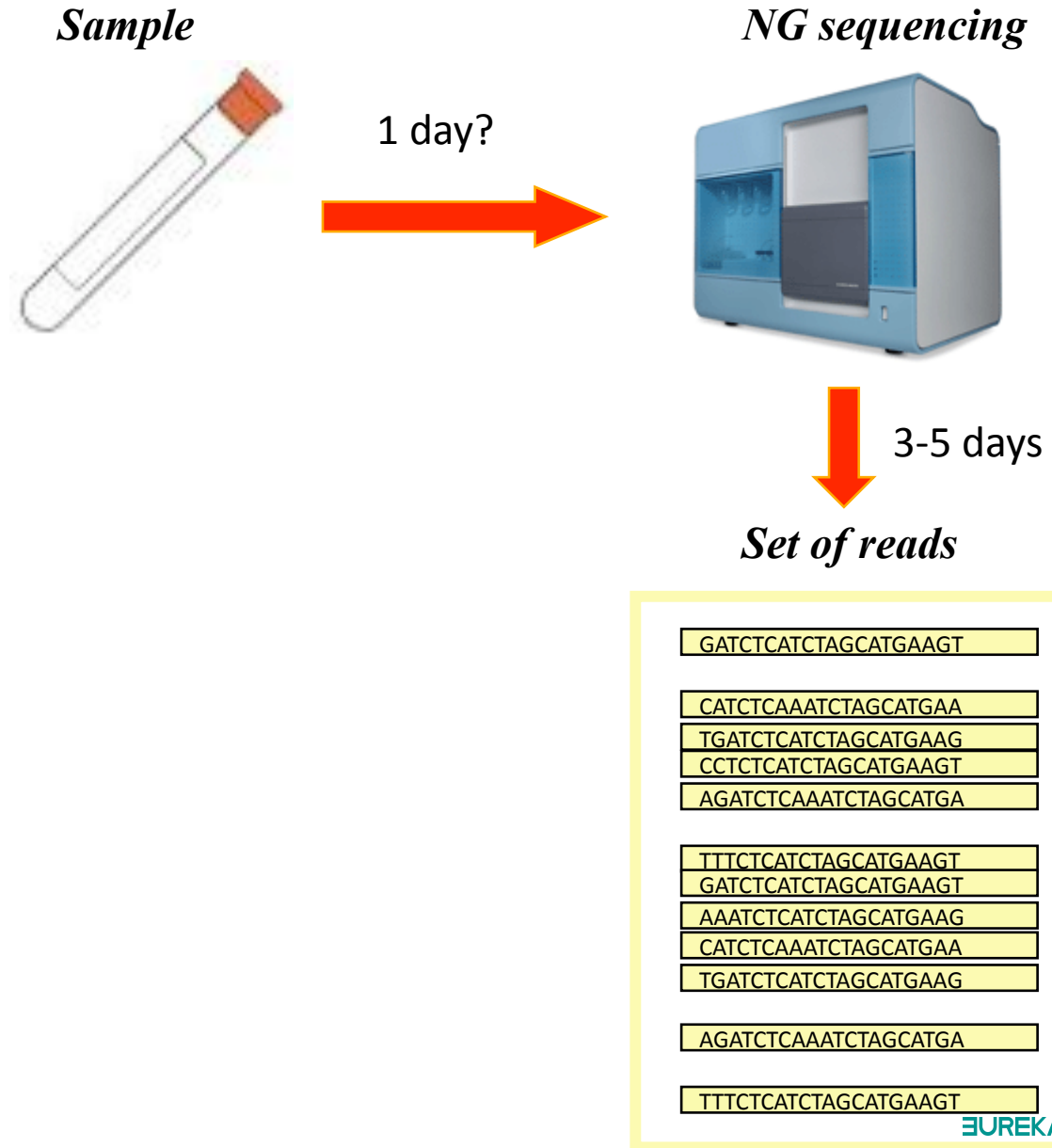
2-5 hours



GATCTCATCTAGCATGAAGT
AAATTCATCTAGCATGAAG
CATCTCAAATCTAGCATGAA
TGATTCATCTAGCATGAAG
CCTTCATCTAGCATGAAGT
AGATCTCAAATCTAGCATGA
GATCTCATCTAGCATGAAGT
TTTTCATCTAGCATGAAGT
GATCTCATCTAGCATGAAGT
AAATTCATCTAGCATGAAG
CATCTCAAATCTAGCATGAA
TGATTCATCTAGCATGAAG
CCTTCATCTAGCATGAAGT
AGATCTCAAATCTAGCATGA
GATCTCATCTAGCATGAAGT
TTTTCATCTAGCATGAAGT



Detection of unknown pathogens





Detection of unknown pathogens

Sample



1 day?



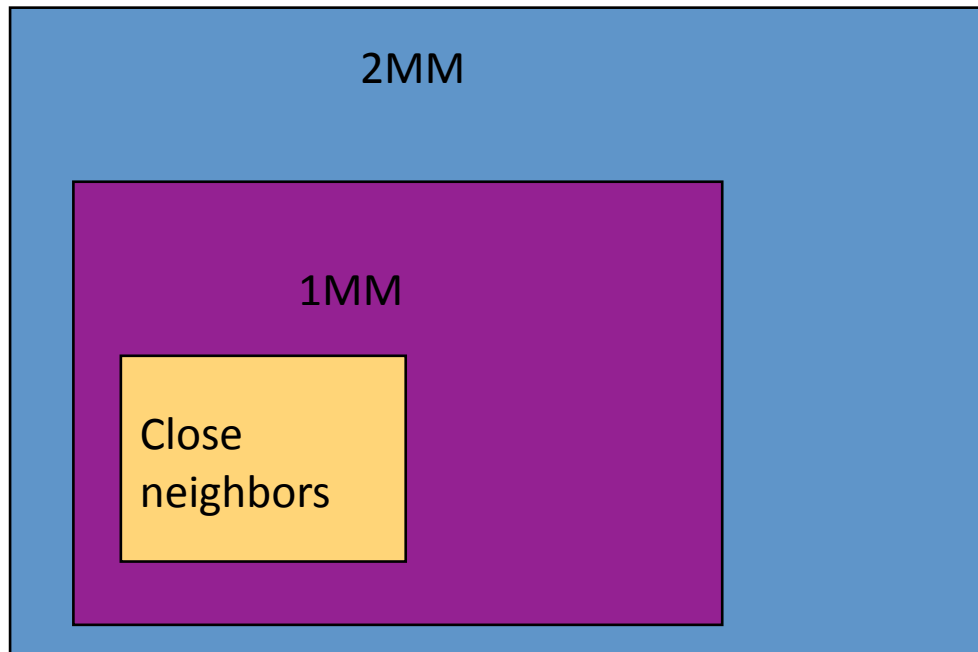
NG sequencing



3-5 days



Set of reads



GATCTCATCTAGCATGAAGT
CATCTCAAATCTAGCATGAA
TGATCTCATCTAGCATGAAG
CCTCTCATCTAGCATGAAGT
AGATCTCAAATCTAGCATGA
TTTCTCATCTAGCATGAAGT
GATCTCATCTAGCATGAAGT
AAATCTCATCTAGCATGAAG
CATCTCAAATCTAGCATGAA
TGATCTCATCTAGCATGAAG
AGATCTCAAATCTAGCATGA
TTTCTCATCTAGCATGAAGT



Detection of unknown pathogens

Sample



1 day?



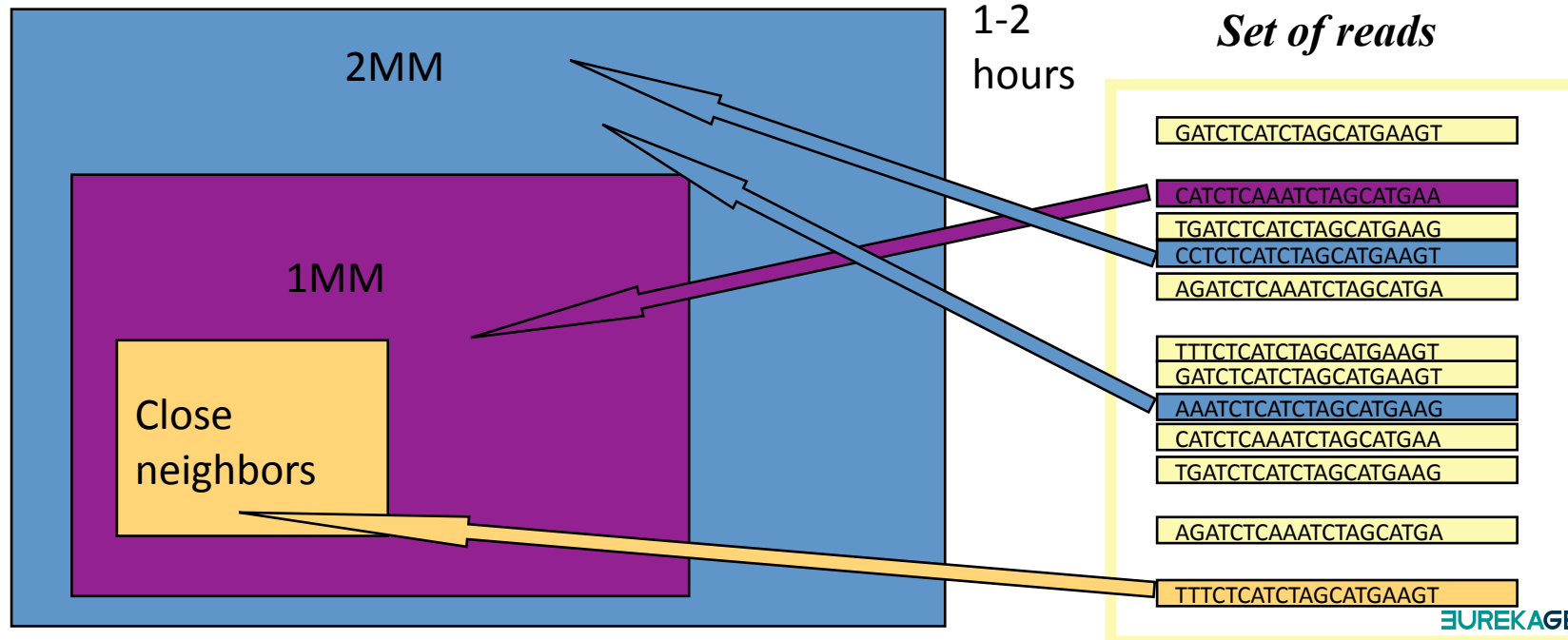
NG sequencing



3-5 days

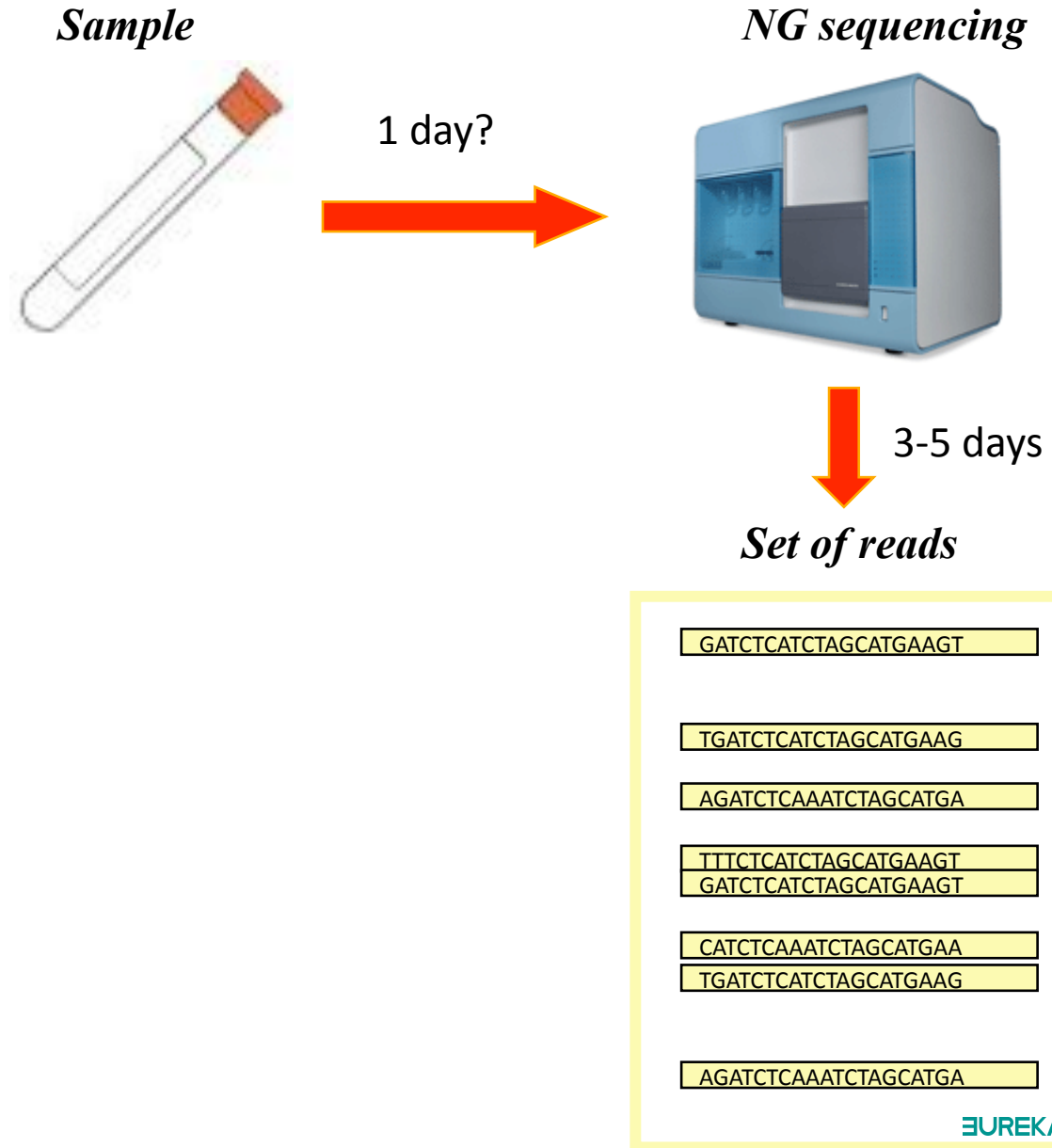


Set of reads





Detection of unknown pathogens





Detection of unknown pathogens

Sample



1 day?



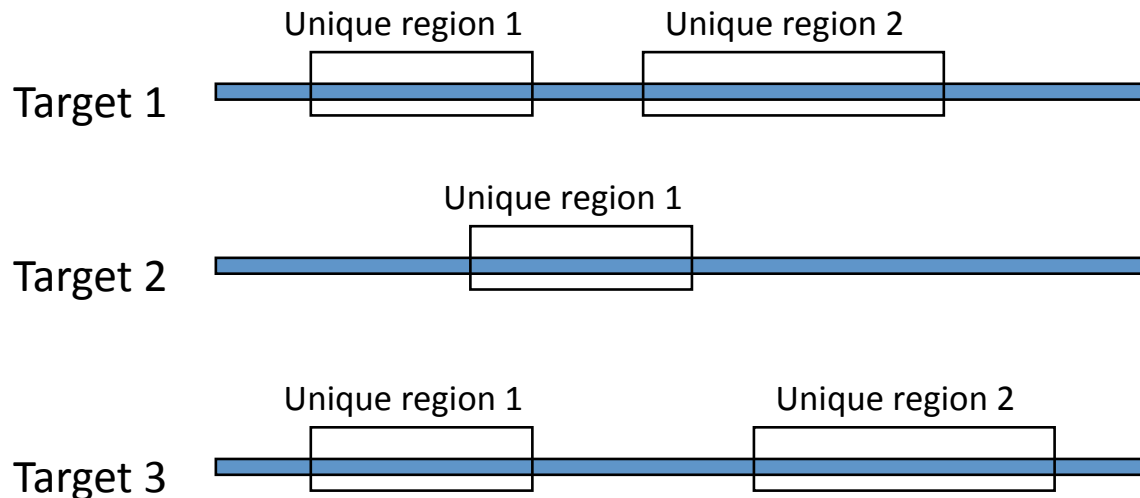
NG sequencing



3-5 days



Set of reads



Set of reads

- GATCTCATCTAGCATGAAGT
- TGATCTCATCTAGCATGAAG
- AGATCTCAAATCTAGCATGA
- TTTCTCATCTAGCATGAAGT
- GATCTCATCTAGCATGAAGT
- CATCTCAAATCTAGCATGAA
- TGATCTCATCTAGCATGAAG
- AGATCTCAAATCTAGCATGA

EUREKAGE NOMICS



Detection of unknown pathogens

Sample



1 day?



NG sequencing

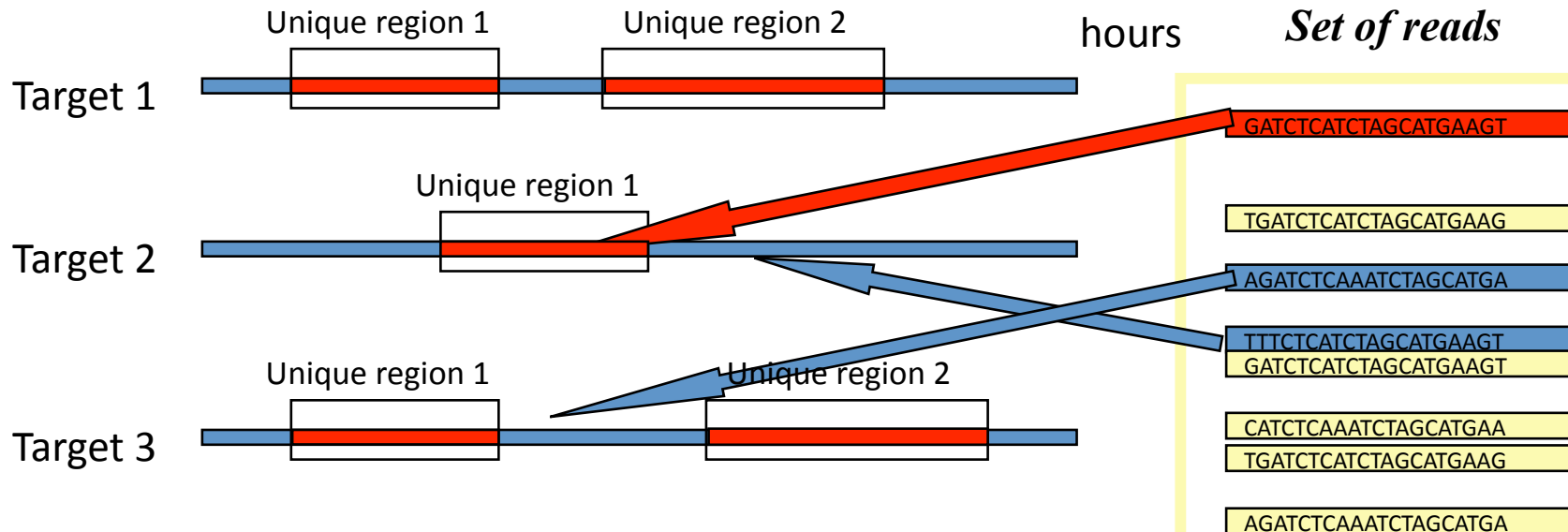


3-5 days



1-2 hours

Set of reads





Detection of unknown pathogens

Sample



1 day?



NG sequencing



3-5 days



Set of reads

TGATCTCATCTAGCATGAAG

GATCTCATCTAGCATGAAGT

CATCTCAAATCTAGCATGAA

TGATCTCATCTAGCATGAAG

AGATCTCAAATCTAGCATGA



Detection of unknown pathogens

Sample



1 day?



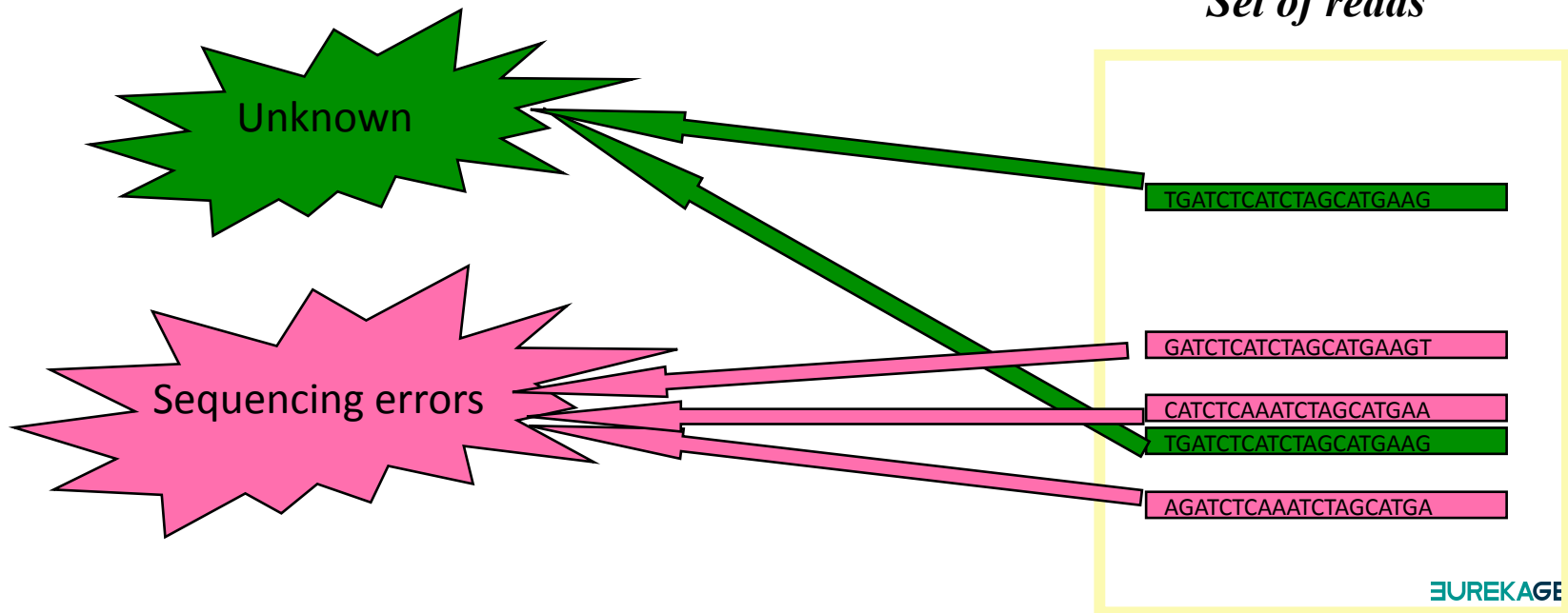
NG sequencing



3-5 days



Set of reads





Examples

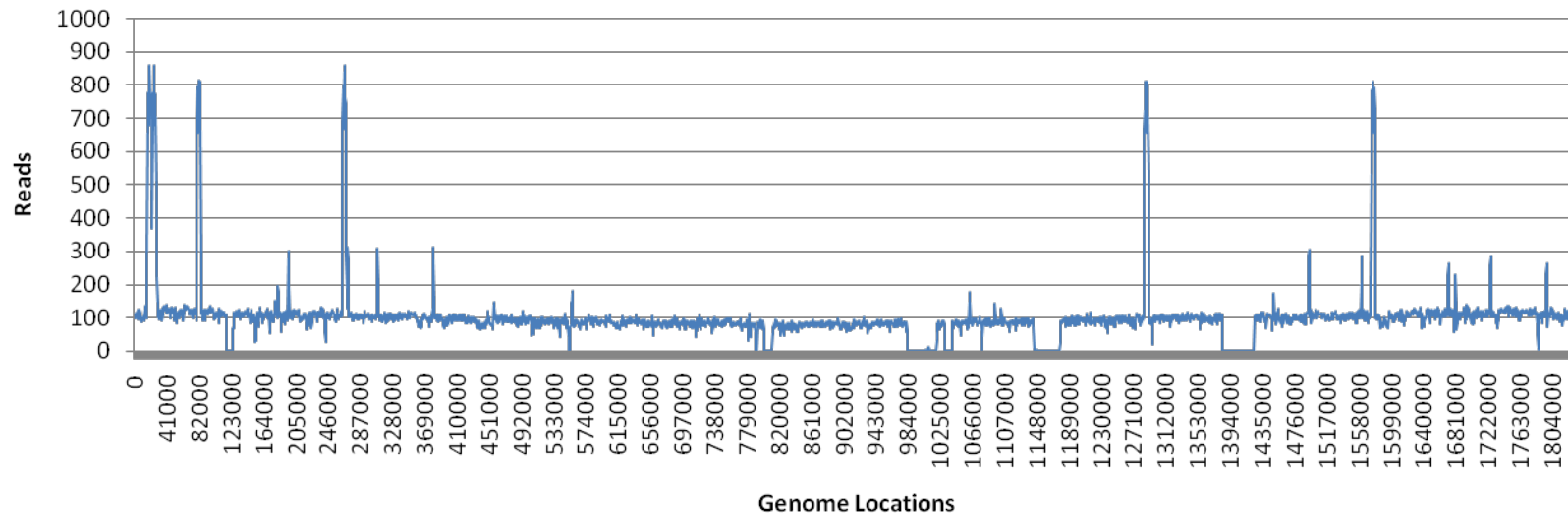


EXAMPLE 1: Bacterial Genome Resequencing

Step 1: Mapping Reads To Reference Genome

- Solexa sequencing of MGAS2109
- 33 base reads
- 8,951,600 reads total
- 160.12x coverage of primary reference genome (MGAS5005)
- Perfect matches: 44.57% 1-mismatches: 9.56%

MGAS2109 Mapping on MGAS5005



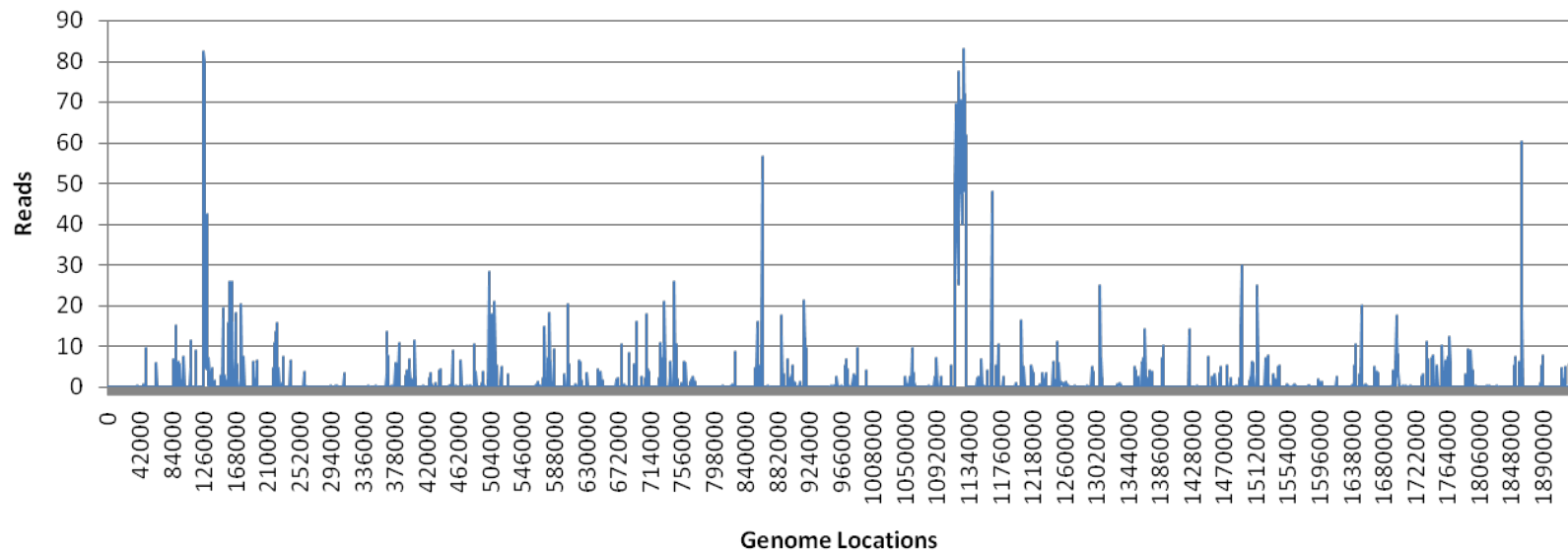


EXAMPLE 1: Bacterial Genome Resequencing

Step 2: Mapping Reads To Reference Relatives

- 4,091,801 (45.87%) reads were ≥ 2 -mismatches from primary reference
- Secondary references: MGAS10270

Remaining Reads Mapped To MGAS10270





EXAMPLE 1: Bacterial Genome Resequencing

Step 3: Assembling Short Reads

Length	Sequence
657	tctaattccatcgctttcatcaaagctacattctcctgtgattcacgcagaactcacaattgcgaatct
428	actatcattagaaagcatcatatggaacaactctattatacggcacaattgattggaatgaaggaca
408	gtgaatcctaacaacaccacaatatgctggaaaacaacgttatcaatTTTaaagaaatccaaa;
394	tgagtataaacagAACCTtaacaaactagctttactaaaaaattaggTctaggcattgcaaaaa;
333	tttcttactttaatatgacggTgatcttgctcaatgaggTtattcagatatttcgatgtacaatgacag
327	tgtggtataggaaaagaaaaaagaaacaaaagcggagatgagatgaaacaaagacttaaccac
306	cctgctaaaaataactaatcgtgacagccaggccctcaactccacTTTTtcttgacgttctcatgcta

Not found

Not found

```
>[ref|NC_008022.1|] D Streptococcus pyogenes MGAS10270, complete genome
Length=1928252
```

Features in this part of subject sequence:

[Fibronectin-binding protein](#)

Score = 712 bits (385), Expect = 0.0
 Identities = 391/394 (99%), Gaps = 0/394 (0%)
 Strand=Plus/Plus

```
>
Length=1928252
```

Features in this part of subject sequence:

[Fibronectin-binding protein](#)

Score = 743 bits (402), Expect = 0.0
 Identities = 406/408 (99%), Gaps = 0/408 (0%)
 Strand=Plus/Plus

```
>[ref|NC_002976.3|] D Staphylococcus epidermidis RP62A, complete genome
Length=2616530
```

Features flanking this part of subject sequence:

[384 bp at 5' side: ADP-ribosylglycohydrolase, putative](#)

[278 bp at 3' side: universal stress protein family, putative](#)

Score = 610 bits (330), Expect = 3e-172
 Identities = 332/333 (99%), Gaps = 0/333 (0%)
 Strand=Plus/Minus

```
>[ref|NC_004350.1|] D Streptococcus mutans UA1
Length=2030921
```

Features in this part of subject sequence:

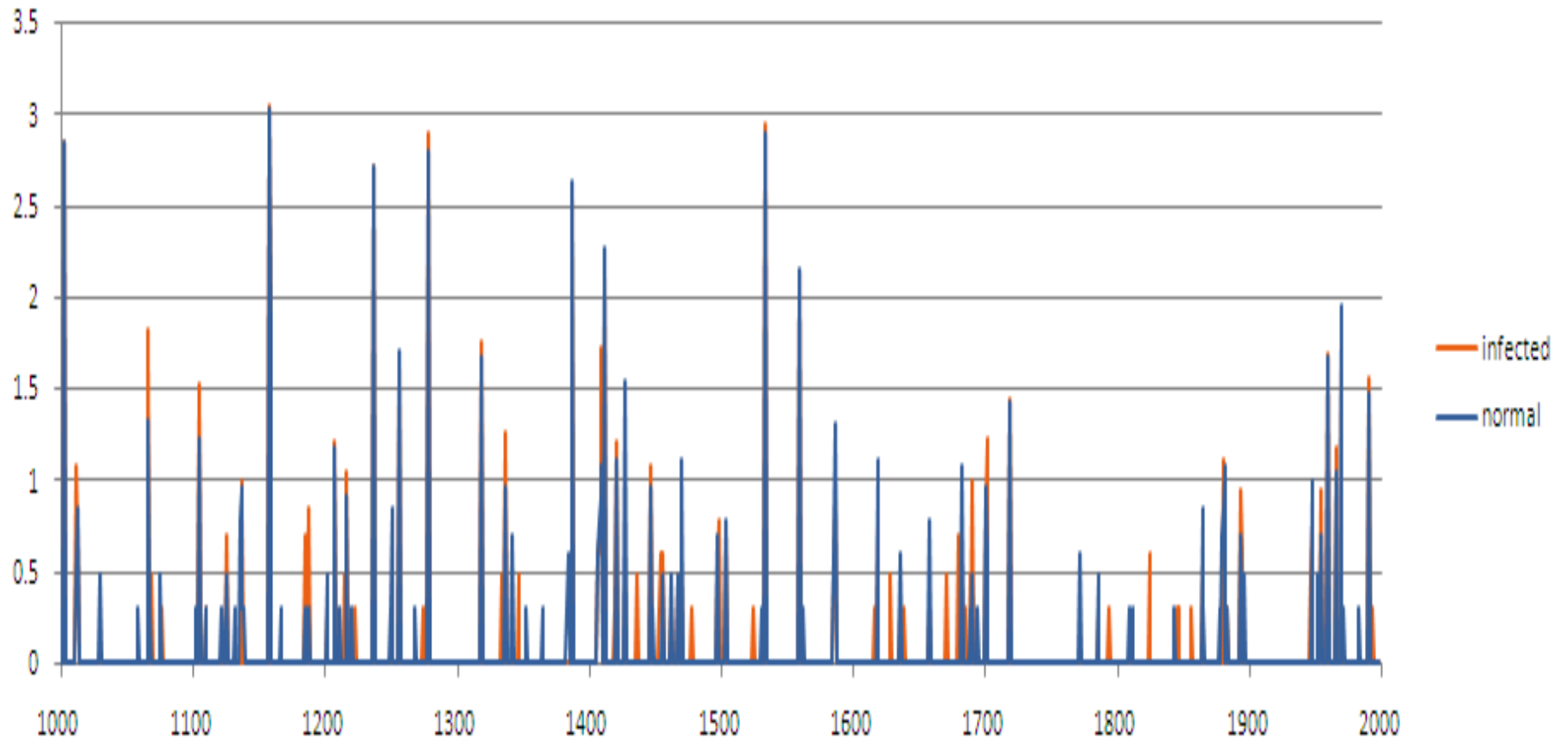
[putative transposase](#)

Score = 134 bits (72), Expect = 1e-28
 Identities = 164/205 (80%), Gaps = 20/205 (9%)
 Strand=Plus/Plus



EXAMPLE 2: Detection of unknown virus

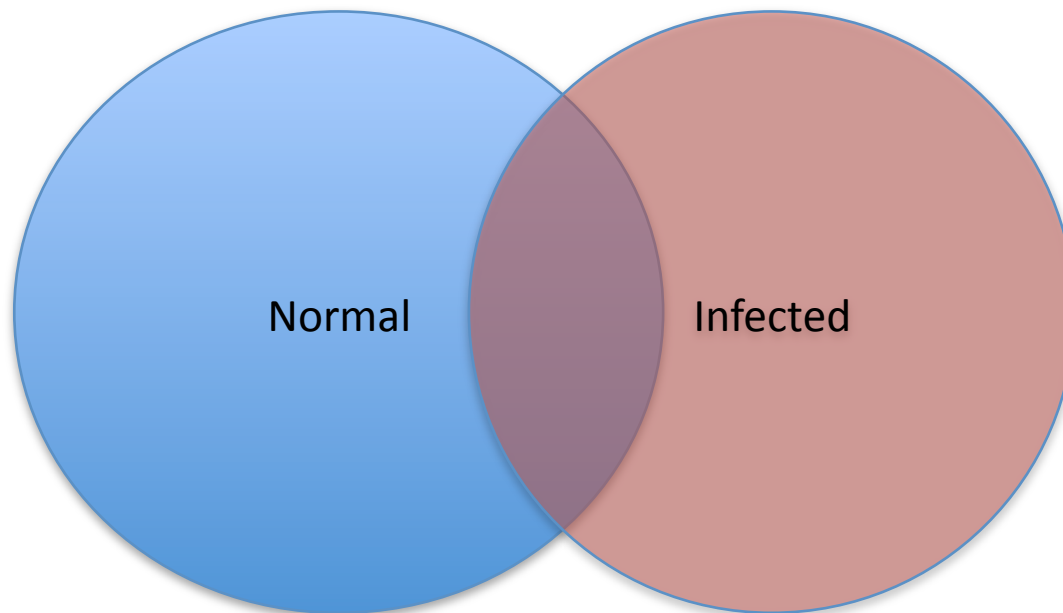
Step 1: Mapping reads (RNA) to the host genome





EXAMPLE 2: Detection of unknown virus

Step 2: Identifying reads UNIQUE to the infected sample





EXAMPLE 2: Detection of unknown virus

Step 3: Assembling reads unique for infected sample and identify (if possible) homologues in GenBank.

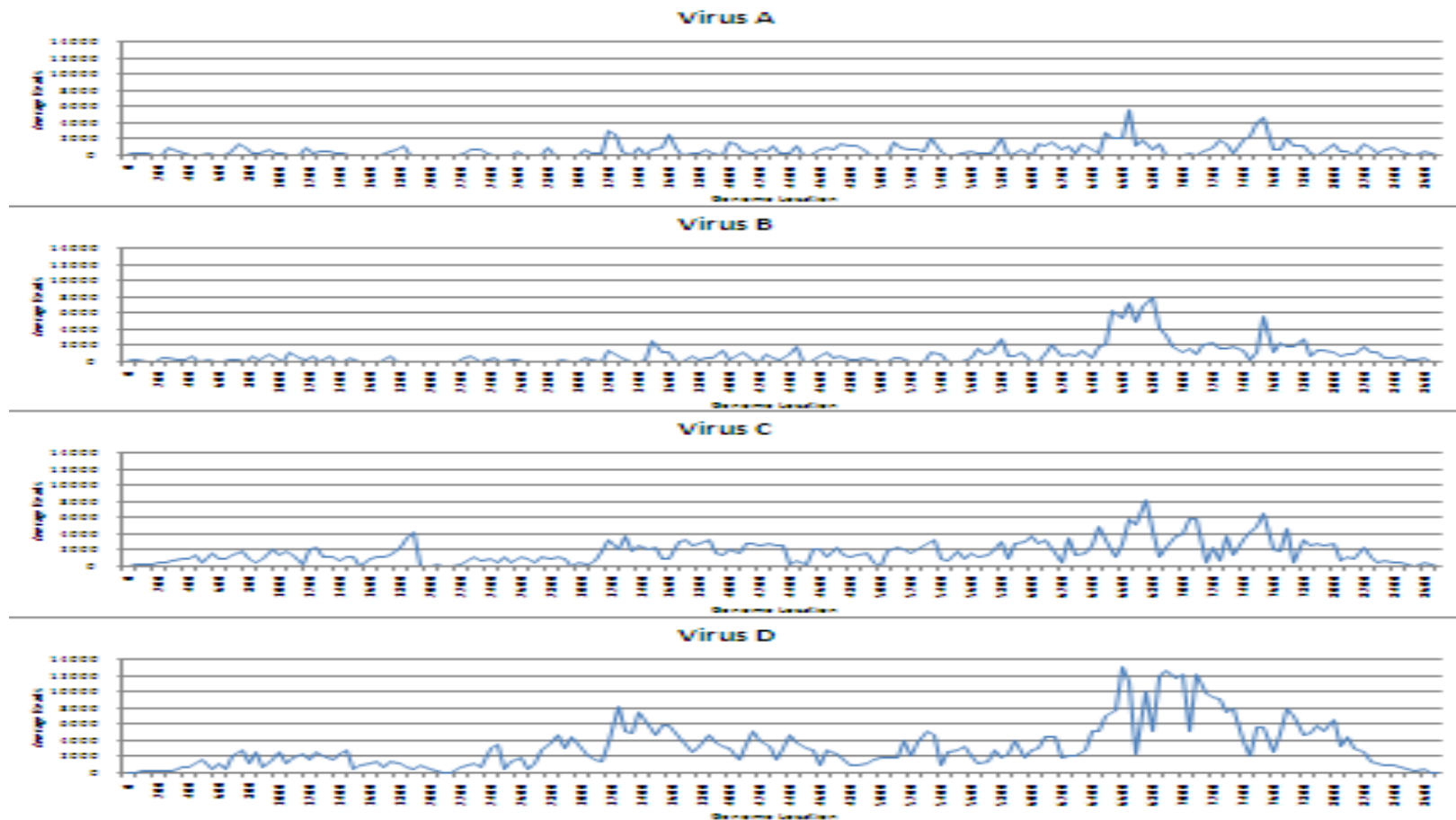
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
AY881626.1	Rupestris stem pitting-associated virus strain SG1, complete genome	740	740	99%	0.0	93%	
AF026278.1	Grapevine Rupestris stem pitting associated virus, complete genome	625	625	100%	7e-176	88%	G
AF057136.1	Rupestris stem pitting associated virus-1, complete genome	625	625	100%	7e-176	88%	G
AY881627.1	Rupestris stem pitting-associated virus strain BS, complete genome	621	621	100%	8e-175	88%	
AY368590.1	Rupestris stem pitting-associated virus strain Syrah, complete genome	486	486	100%	4e-134	82%	
DQ278637.1	Rupestris stem pitting-associated virus isolate Seyve Villard 3160-2 nonfunctional replicase mRNA, partial sequence	271	271	32%	2e-69	97%	
DQ278650.1	Rupestris stem pitting-associated virus isolate Seyval 19 replicase mRNA, partial cds	262	262	32%	9e-67	96%	
DQ278640.1	Rupestris stem pitting-associated virus isolate Ravat 34-6-2 replicase mRNA, partial cds	257	257	32%	4e-65	95%	
DQ278626.1	Rupestris stem pitting-associated virus isolate Trebbiano 12-5 nonfunctional replicase mRNA, partial sequence	257	257	32%	4e-65	95%	
DQ278635.1	Rupestris stem pitting-associated virus isolate Pinot Noir 1 replicase mRNA, partial cds	255	255	32%	1e-64	95%	

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
AY881627.1	Rupestris stem pitting-associated virus strain BS, complete genome	1038	1038	100%	0.0	98%	
AB277787.1	Rupestris stem pitting-associated virus genes for RNA-dependent RNA polymerase, 24.4 kDa hypothetical protein, 13.8 kDa hypothetical protein, 8.4 kDa hypothetical protein, coat protein, partial and complete cds, strain: Hm1	1034	1034	99%	0.0	98%	
BF028294.1	Rupestris stem pitting-associated virus isolate RSP47-4 replicase gene, partial cds; triple gene block protein 1, triple gene block protein 2, and triple gene block protein 3 genes, complete cds; and capsid protein gene, partial cds	917	917	100%	0.0	91%	
AB277788.1	Rupestris stem pitting-associated virus genes for RNA-dependent RNA polymerase, 24.4 kDa hypothetical protein, 13.8 kDa hypothetical protein, 8.4 kDa hypothetical protein, coat protein, partial and complete cds, strain: Hm1	899	899	100%	0.0	90%	
AF057136.1	Rupestris stem pitting associated virus-1, complete genome	887	887	99%	2e-188	79%	G
AB277789.1	Rupestris stem pitting-associated virus genes for RNA-dependent RNA polymerase, 24.4 kDa hypothetical protein, 13.8 kDa hypothetical protein, 8.4 kDa hypothetical protein, coat protein, partial and complete cds, strain: Hm3	888	888	99%	1e-188	78%	G
AF026278.1	Grapevine Rupestris stem pitting associated virus, complete genome	854	854	99%	1e-184	78%	G
AY368590.1	Rupestris stem pitting-associated virus strain Syrah, complete genome	848	848	99%	7e-182	78%	
AB277784.1	Rupestris stem pitting-associated virus genes for RNA-dependent RNA polymerase, 24.4 kDa hypothetical protein, 13.8 kDa hypothetical protein, 8.4 kDa hypothetical protein, coat protein, partial and complete cds, strain: CBS	838	838	99%	4e-149	78%	



EXAMPLE 2: Detection of unknown virus

Step 4: Mapping all reads to identified homologues





Didier Perez

didier@eurekagenomics.com

415-269-0666