Practical Insights into Bacterial Genomic Annotation

Julio Diaz Caballero
University of Oxford

wellcome connecting science

ACORN

Centre for Genomic Pathogen Surveillance

BIG DATA INSTITUTE

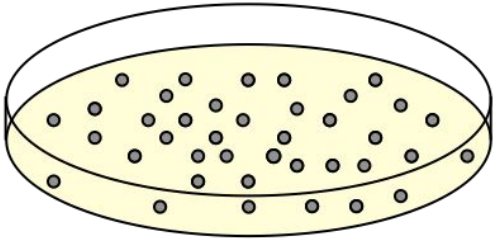
UNIVERSITY OF OXFORD

Genome annotation is the process of **identifying genes and other functional elements** within a DNA sequence and attaching information about their possible functions. It transforms a **sequence of nucleotides into a functional map of genetic capabilities**.

Why Annotate Bacterial Genomes?

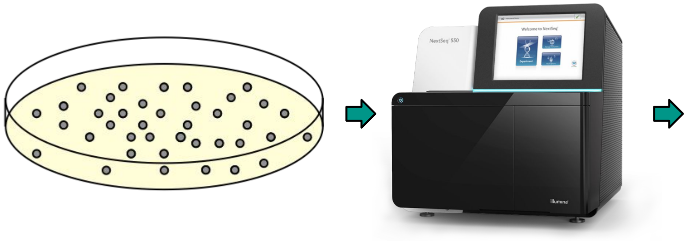
- Better describe the functional capability of a bacteria
- Identify risk (AMR, virulence, etc) / beneficial (biofuels, etc) factors
- Typing bacteria to find similar strains (Surveillance)
- Identify potential adaptation
- Comparative Genomics
- Microbiome studies

From Bacterial Samples to Genetic Content



```
> chromosome
TAACCGGTAAAACACTTTGCTCGTGCACATAGG
CGTGACACCTAAGACTGGAACAAGCTCAGAAG
AGTAGTAGGCGAGCATTTTTTGACCGAGTCCG
CTCCTTTCTAACTCACTGACTTCTCGCGGGCCG
TATCCTGCACGCTCAACAGCCAGCGGTGTCCC
GTTACCCTTCAAGCTCATCTTCCTCGAGGTCTG
TTGTAGTACCACACGCCTCTCCCGGCATTAGC
CCGCACTCCTCGACGGGACATTATGTGCCTTC
AGTTCCCGATCTCGGTCGCGCCCAGCCGGAAT
CCCTTAGACACCAGGGCCGCGTGAGCGAGAA
GCGGGGGGAGAACTTTATAGGGCTGTGGCTCA
TACAATAGGGTAAGGTTTCACCACATTTCTTCA
CTTCAGAAGCGACGCCTCCATTTTGCCCTCACC
CACGGTATAAGACGAAAGCCTAAGGCAACCCC
GGGGTTAGACGCGGTCCTTCTCTCTACT
```

From Bacterial Samples to Genetic Content



```
> chromosome
TAACCGGTAAAACACTTTGCTCGTGCACATAGG
CGTGACACCTAAGACTGGAACAAGCTCAGAAG
AGTAGTAGGCGAGCATTTTTTGACCGAGTCCG
CTCCTTTCTAACTCACTGACTTCTCGCGGGCCG
TATCCTGCACGCTCAACAGCCAGCGGTGTCCC
GTTACCCTTCAAGCTCATCTTCCTCGAGGTCTG
TTGTAGTACCACACGCCTCTCCCGGCATTAGC
CCGCACTCCTCGACGGGACATTATGTGCCTTC
AGTTCCCGATCTCGGTCGCGCCCAGCCGGAAT
CCCTTAGACACCAGGGCCGCGTGAGCGAGAA
GCGGGGGGAGAACTTTATAGGGCTGTGGCTCA
TACAATAGGGTAAGGTTTCACCACATTTCTTCA
CTTCAGAAGCGACGCCTCCATTTTGCCCTCACC
CACGGTATAAGACGAAAGCCTAAGGCAACCCC
GGGGTTAGACGCGGTCCTTCTCTCTACT
```

GENOME ANNOTATION

Genetic Potential

CGPS

What is Genome Annotation?

```
> chromosome
TAACCGGTAAAACACTTTGCTCGTGCACATAGG
CGTGACACCTAAGACTGGAACAAGCTCAGAAG
AGTAGTAGGCGAGCATTTTTTGACCGAGTCCG
CTCCTTTCTAACTCACTGACTTCTCGCGGGCCG
TATCCTGCACGCTCAACAGCCAGCGGTGTCCC
GTTACCCTTCAAGCTCATCTTCCTCGAGGTCTG
TTGTAGTACCACACGCCTCTCCCGGCATTAGC
CCGCACTCCTCGACGGGACATTATGTGCCTTC
AGTTCCCGATCTCGGTCGCGCCCAGCCGGAAT
CCCTTAGACACCAGGGCCGCGTGAGCGAGAA
GCGGGGGGAGAACTTTATAGGGCTGTGGCTCA
TACAATAGGGTAAGGTTTCACCACATTTCTTCA
CTTCAGAAGCGACGCCTCCATTTTGCCCTCACC
CACGGTATAAGACGAAAGCCTAAGGCAACCCC
GGGGTTAGACGCGGTCCTTCTCTCTACT
```

Genetic Potential:

- Gene prediction
- Regulatory genes
- Structural motifs
- Mobile Genetic Elements

All we have are As, Ts, Gs, and Cs mixed up

```
> chromosome
TAACCGGTAAAACACTTTGCTCGTGCACATAGGCGTGACACCTAAGACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTT
TTGACCGAGTCCGCTCCTTTCTAACTCACTGACTTCTCGCGGGCCGTATCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACC
CTTCAAGCTCATCTTCCTCGAGGTCTGTTGTAGTACCACACGCCTCTCCCGGCATTAGCCCGCACTCCTCGACGGGACATTATG
TGCCTTCAGTTCCCGATCTCGGTCGCGCCCAGCCGGAATCCCTTAGACACCAGGGCCGCGTGAGCGAGAAGCGGGGGGAGAA
CTTTATAGGGCTGTGGCTCATACAATAGGGTAAGGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCAC
CCACGGTATAAGACGAAAGCCTAAGGCAACCCCGGGGTTAGACGCGGTCCTTCTCTCTACTTAACCGGTAAAACACTTTGCTCG
TGCACATAGGCGTGACACCTAAGACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTTTTGACCGAGTCCGCTCCTTTCTA
ACTCACTGACTTCTCGCGGGCCGTATCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACCCTTCAAGCTCATCTTCCTCGAGG
TCTGTTGTAGTACCACACGCCTCTCCCGGCATTAGCCCGCACTCCTCGACGGGACATTATGTGCCTTCAGTTCCCGATCTCGGT
CGCGCCCAGCCGGAATCCCTTAGACACCAGGGCCGCGTGAGCGAGAAGCGGGGGGAGAACTTTATAGGGCTGTGGCTCATAC
AATAGGGTAAGGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCACCCACGGTATAAGACGAAAGCCTA
AGGCAACCCCGGGGTTAGACGCGGTCCTTCTCTCTACTTAACCGGTAAAACACTTTGCTCGTGCACATAGGCGTGACACCTAAG
ACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTTTTGACCGAGTCCGCTCCTTTCTAACTCACTGACTTCTCGCGGGCCG
TATCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACCCTTCAAGCTCATCTTCCTCGAGGTCTGTTGTAGTACCACACGCCTC
TCCCGGCATTAGCCCGCACTCCTCGACGGGACATTATGTGCCTTCAGTTCCCGATCTCGGTCGCGCCCAGCCG
```

All we have are As, Ts, Gs, and Cs mixed up

```
> chromosome
TAACCGGTAAAACACTTTGCTCGTGCACATAGGCGTGACACCTAAGACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTT
TTGACCGAGTCCGCTCCTTTCTAACTCACTGACTTCTCGCGGGCCGTATCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACC
CTTCAAGCTCATCTTCCTCGAGGTCTGTTGTAGTACCACACGCCTCTCCCGGCATTAGCCCGCACTCCTCGACGGGACATTATG
TGCCTTCAGTTCCCGATCTCGGTCGCGCCCAGCCGGAATCCCTTAGACACCAGGGCCGCGTGAGCGAGAAGCGGGGGGAGAA
CTTTATAGGGCTGTGGCTCATACAATAGGGTAAGGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCAC
CCACGGTATAAGACGAAAGCCTAAGGCAACCCCGGGGTTAGACGCGGTCCTTCTCTCTACTTAACCGGTAAAACACTTTGCTCG
TGCACATAGGCGTGACACCTAAGACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTTTTGACCGAGTCCGCTCCTTTCTA
ACTCACTGACTTCTCGCGGGCCGTATCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACCCTTCAAGCTCATCTTCCTCGAGG
TCTGTTGTAGTACCACACGCCTCTCCCGGCATTAGCCCGCACTCCTCGACGGGACATTATGTGCCTTCAGTTCCCGATCTCGGT
CGCGCCCAGCCGGAATCCCTTAGACACCAGGGCCGCGTGAGCGAGAAGCGGGGGGAGAACTTTATAGGGCTGTGGCTCATAC
AATAGGGTAAGGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCACCCACGGTATAAGACGAAAGCCTA
AGGCAACCCCGGGGTTAGACGCGGTCCTTCTCTCTACTTAACCGGTAAAACACTTTGCTCGTGCACATAGGCGTGACACCTAAG
ACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTTTTGACCGAGTCCGCTCCTTTCTAACTCACTGACTTCTCGCGGGCCG
TATCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACCCTTCAAGCTCATCTTCCTCGAGGTCTGTTGTAGTACCACACGCCTC
TCCCGGCATTAGCCCGCACTCCTCGACGGGACATTATGTGCCTTCAGTTCCCGATCTCGGTCGCGCCCAGCCG
```

How do we annotate a genome?

The Annotation Process

- Step 1: Input and Quality Check
- Step 2: Gene Prediction
- Step 3: Functional Prediction
- Step 4: Structural RNA Prediction
- Step 5: Validation and Refinement

The Annotation Process

Step 1: Input and Quality Check

➔ Assembled Genome in FASTA format

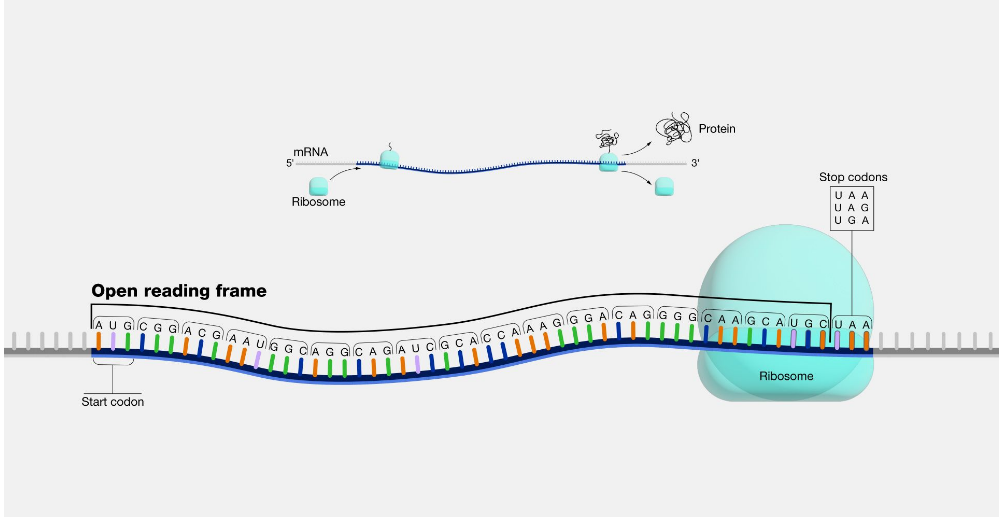
➔ The more contiguous the better

➔ Long read assemblies may resolve certain annotations better

The Annotation Process

Step 2: Gene Prediction

Open reading frames (ORF)


mRNA
5'
3'
Ribosome
Protein
Stop codons
U A A
U A G
U G A
Open reading frame
A U G C G G A C G A A U G G C A G G C A G A U C G C A C C A A A G G G A C A G G G G C A A G C A U G C U A A
Start codon
Ribosome

https://www.genome.gov/genetics-glossary/Open-Reading-Frame

The Annotation Process

Step 2: Gene Prediction

Ab-initio prediction

```
> chromosome
TAACCGATGTAACACTTTGCTCGTGCACATAGGCGTGACACCTAAGACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTTTTGA

CCGAGTCCGCTCCTTTCTAACTCACTGACTTCTCGCGGGCCATGTCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACCCTTCAAG

CTCATCTTCCTCGAGGTCTGTTGTAGTACCACACGCCTCTCCCGGCATATGCCCGCACTCCTCGACGGGACATTATGTGCCTTCAGTT

CCCGATCTCGGTCGCGCCCAGCCGGAATCCCTATGGCACCAGGGCCGCGTGAGCGAGAAGCGGGGGAGAACTTTATAGGGCTGT

GGCTCATACAATAGGGTAAGGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCACCCACGGTATAAGACGAA

AGCCTAAGGCAACCCCGGGGTATGGCGCGGTCCTTCTCTCTACTTAACCGGTAAAACACTTTGCTCGTGCACATAGGCGTGACACCT
```

The Annotation Process

Start Codon	**ATG**

Step 2: Gene Prediction

Ab-initio prediction of open reading frames (ORF)

> chromosome

TAACCG**ATG**TAACACTTTGCTCGTGCACATAGGCGTGACACCTAAGACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTTTTGA

CCGAGTCCGCTCCTTTCTAACTCACTGACTTCTCGCGGGCC**ATG**TCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACCCTTCAAG

CTCATCTTCCTCGAGGTCTGTTGTAGTACCACACGCCTCTCCCGGCAT**ATG**CCCGCACTCCTCGACGGGACATT**ATG**TGCCTTCAGTT

CCCGATCTCGGTCGCGCCCAGCCGGAATCCCT**ATG**GCACCAGGGCCGCGTGAGCGAGAAGCGGGGGGAGAACTTTATAGGGCTGT

GGCTCATACAATAGGGTAAGGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCACCCACGGTATAAGACGAA

AGCCTAAGGCAACCCCGGGGTATGGCGCGGTCCTTCTCTCTACTTAACCGGTAAAACACTTTGCTCGTGCACATAGGCGTGACACCT

The Annotation Process

Step 2: Gene Prediction

E.g. Ab-initio prediction of open reading frames (ORF)

Start Codon	**ATG**
Stop Codon	**TAG** **TGA** **TAA**

> chromosome

TAACCG**ATG**TAACACTTTGCTCGTGCACA**TAG**GCG**TGA**CACC**TAA**GACTGGAACAAGCTCAGAAGAG**TAGTAG**GCGAGCATTTTT**TGA**

CCGAGTCCGCTCCTTTC**TAA**CTCAC**TGA**CTTCTCGCGGGCC**ATG**TCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACCCTTCAAG

CTCATCTTCCTCGAGGTCTGTTG**TAG**TACCACACGCCTCTCCCGGCAT**ATG**CCCGCACTCCTCGACGGGACATT**ATG**TGCCTTCAGTT

CCCGATCTCGGTCGCGCCCAGCCGGAATCCCT**ATG**GCACCAGGGCCGCG**TGA**GCGAGAAGCGGGGGGAGAACTTTA**TAG**GGCTGT

GGCTCATACAA**TAG**GG**TAA**GGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCACCCACGGTA**TAA**GACGAA

AGCC**TAA**GGCAACCCCGGGGTATGGCGCGGTCCTTCTCTCTACT**TAA**CCGG**TAA**AACACTTTGCTCGTGCACA**TAG**GCG**TGA**CACCT

The Annotation Process

Start Codon	**ATG**
Stop Codon	**TAG** **TGA** **TAA**

Step 2: Gene Prediction

Ab-initio prediction of open reading frames (ORF)

> chromosome

TAACCG**ATG**TAACACTTTGCTCGTGCACA**TAG**GCG**TGA**CACC**TAA**GACTGGAACAAGCTCAGAAGAG**TAGTAG**GCGAGCATTTTT**TG**
A

ORF1

CCGAGTCCGCTCCTTTC**TAA**CTCAC**TGA**CTTCTCGCGGGCC**ATG**TCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACCCTTCAAG

ORF2

CTCATCTTCCTCGAGGTCTGTTG**TAG**TACCACACGCCTCTCCCGGCAT**ATG**CCCGCACTCCTCGACGGGACATT**ATG**TGCCTTCAGTT

CCCGATCTCGGTCGCGCCCAGCCGGAATCCCT**ATG**GCACCAGGGCCGCG**TGA**GCGAGAAGCGGGGGAGAACTTTA**TAG**GGCTGT

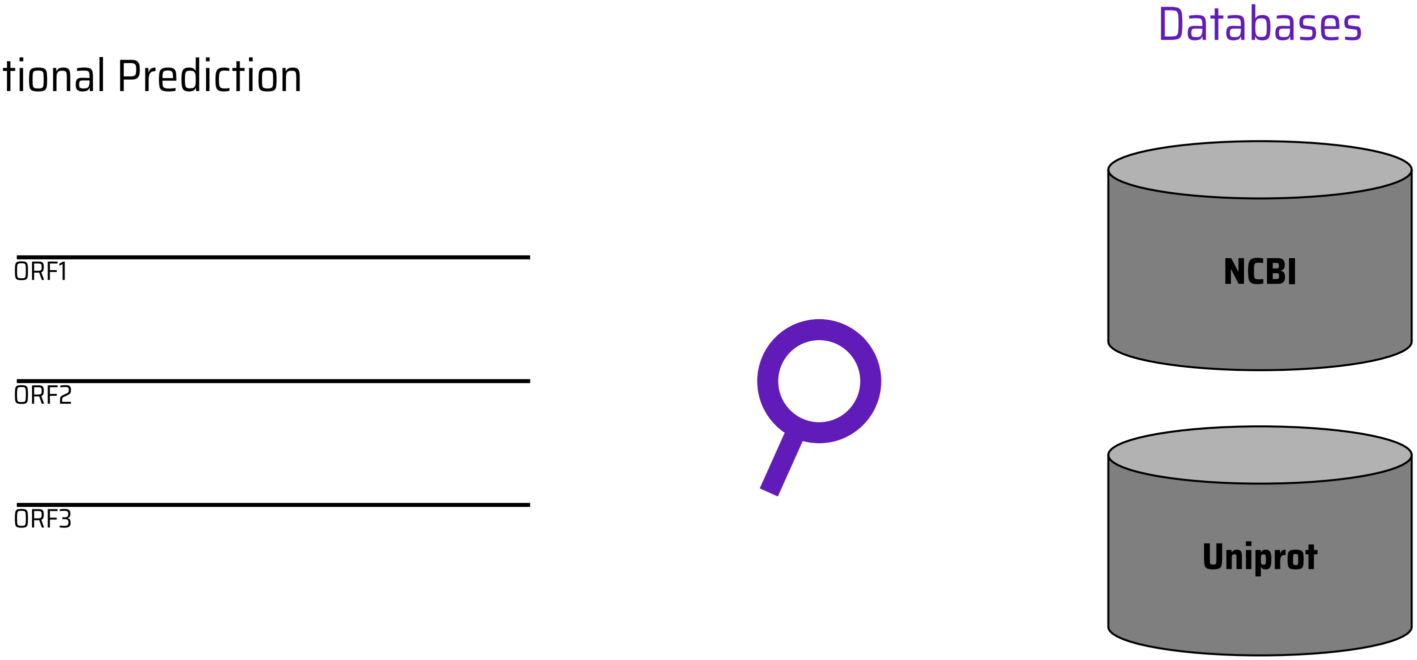
ORF3

GGCTCATACAA**TAG**GG**TAA**GGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCACCCACGGTA**TAA**GACGAA

AGCC**TAA**GGCAACCCCGGGGTATGGCGCGGTCCTTCTCTCTACT**TAA**CCGG**TAA**AACACTTTGCTCGTGCACA**TAG**GCG**TGA**CACCT

The Annotation Process

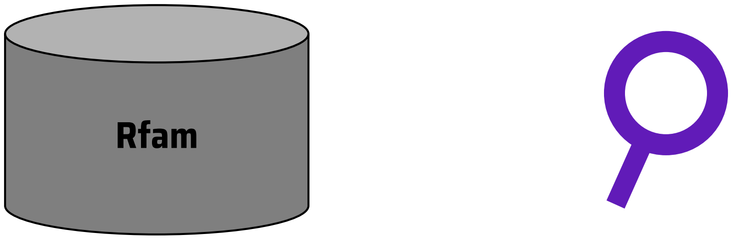
Step 3: Functional Prediction



The Annotation Process

Step 4: Structural RNA Prediction
(tRNAs, rRNAs, other non-coding RNA)

Databases



```
> chromosome
UAACCGAUGUAACACUUUGCUCGUGCACAUAGGCGUGACACCUAAGACUGGAACAAGCCACAAGC

CCGAGUCCGCUCCUUUCUAACUCACUGACUUCUCGCGGGCCAUGUCCUGCACGCUCAACACAAGC

CUCAUCUUCCUCGAGGUCUGUUGUAGUACCACACGCCUCUCCCGGCAUAUGCCCGCACUACAAGC

CCCGAUCUCGGUCGCGCCCAGCCGGAAUCCCUAUGGCACCAGGGCCGCGUGAGCGAGAACAAGC

GGCUCAUACAAUAGGGUAAGGUUUCACCACAUUUCUUCACUUCAGAAGCGACGCCUCCAACAAGC

AGCCUAAGGCAACCCCGGGGUAUGGCGCGGUCCUUCUCUCUACUUAACCGGUAAAACACACAAGC
```

The Annotation Process

Step 4: Structural RNA Prediction
(tRNAs, rRNAs, other non-coding RNA)

> chromosome

UAACCGAUGUAACACUUUGCUCGUGCACAUAGGCGUGACACCUAAGACUGGAACAAGCUCAGAAGAGUAGUAGGCGAGCAUUUUUUGA

ORF1

CCGAGUCCGCUCCUUUCUAACUCACUGACUUCUCGCGGGCCAUGUCCUGCACGCUCAACAGCCAGCGGUGUCCCGUUACCCUUCAAG

ORF2

CUCAUCUUCCUCGAGGUCUGUUGUAGUACCACACGCCUCUCCCGGCAUAUGCCCGCACUCCUCGACGGGACAUUAUGUGCCUUCAGUU

CCCGAUCUCGGUCGCGCCCAGCCGGAAUCCCUAUGGCACCAGGGCCGCGUGAGCGAGAAGCGGGGGGAGAACUUUAUAGGGCUGU

ORF3

GGCUCAUACAAUAGGGUAAGGUUUCACCACAUUUCUUCACUUCAGAAGCGACGCCUCCAUUUUGCCCUCACCCACGGUAUAAGACGAA

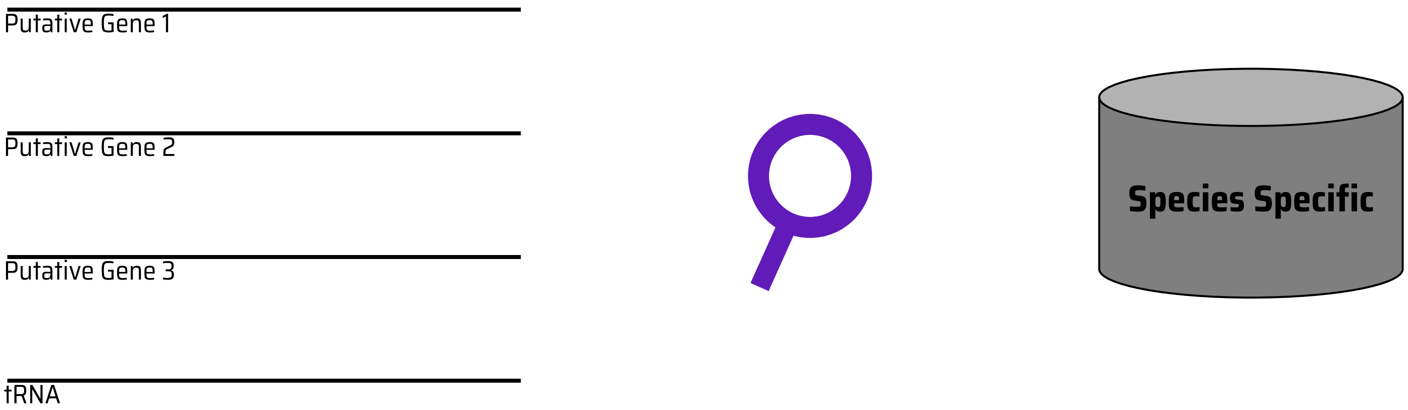
AGCCUAAGGCAACCCCGGGGUAUGGCGCGGUCCUUCUCUCUACUUAACCGGUAAAACACUUUGCUCGUGCACAUAGGCGUGACACCU

tRNA

The Annotation Process

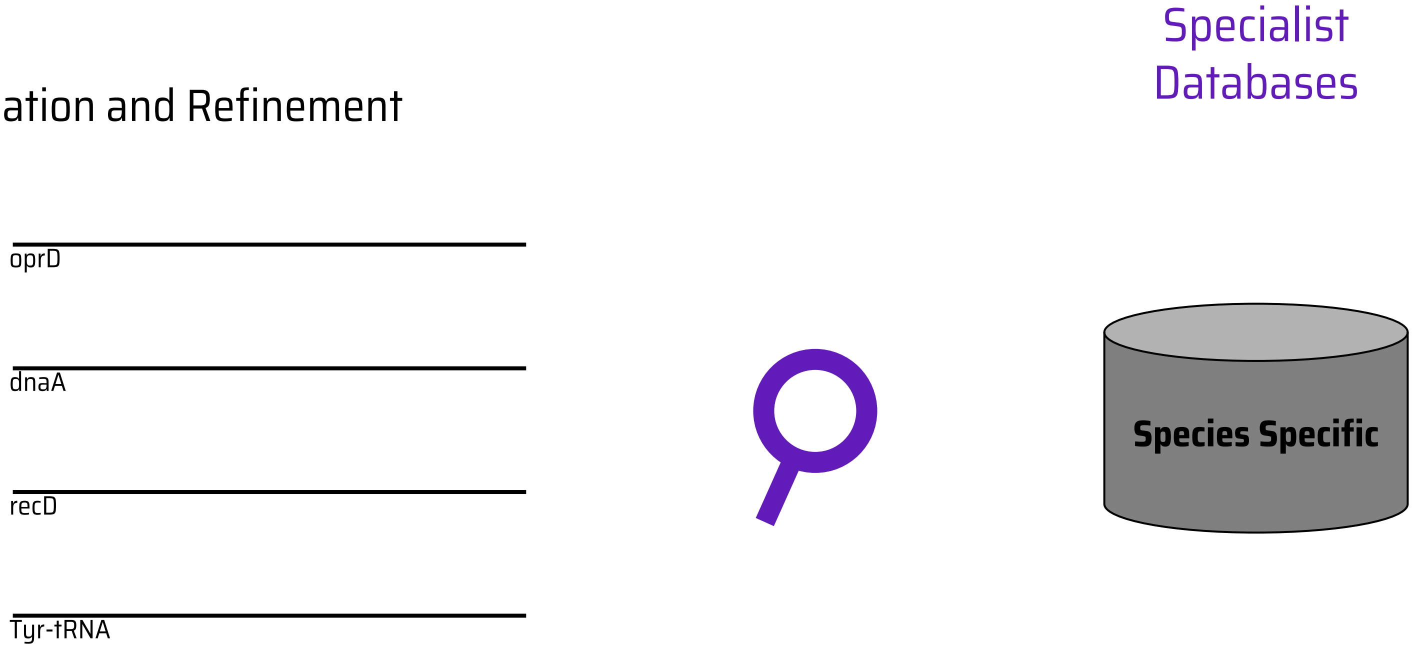
Step 5: Validation and Refinement

Specialist Databases

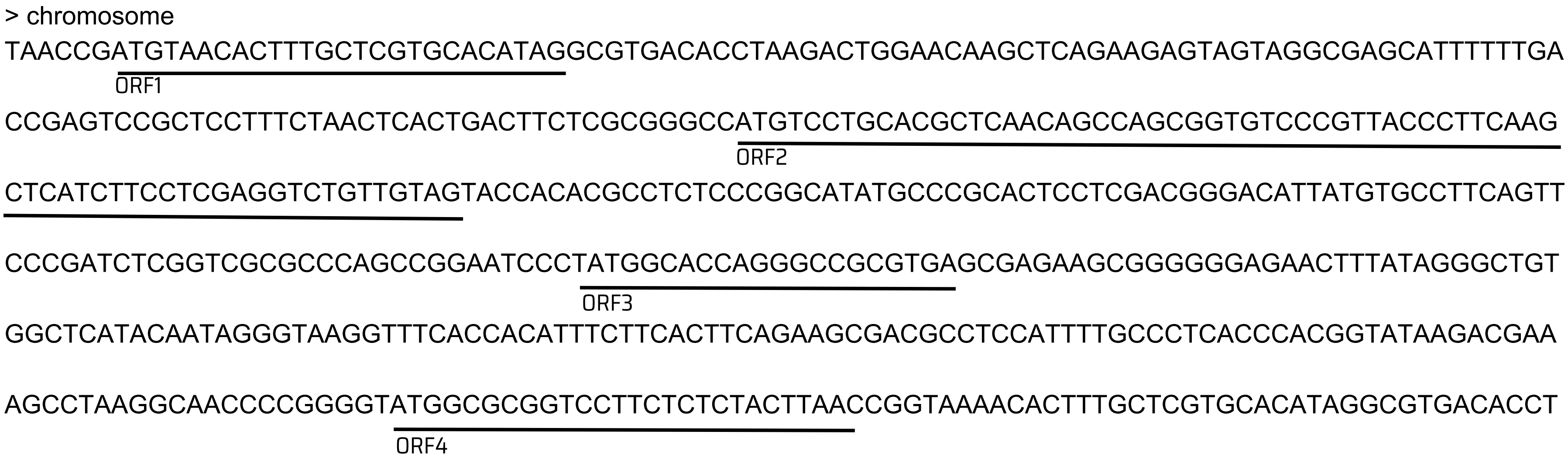


The Annotation Process

Step 5: Validation and Refinement

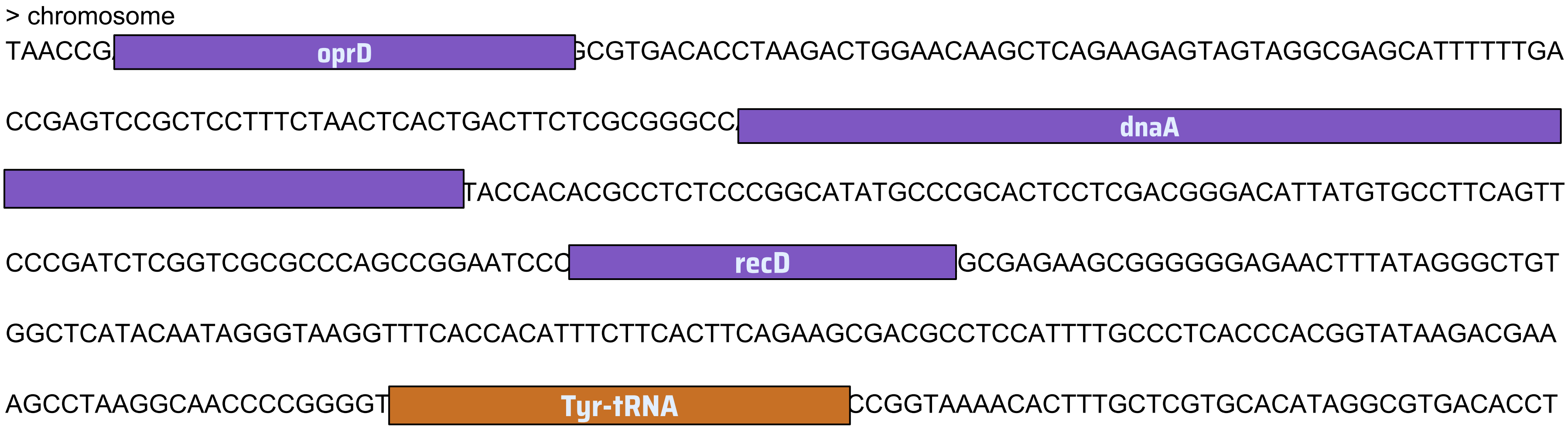


Annotated Genome



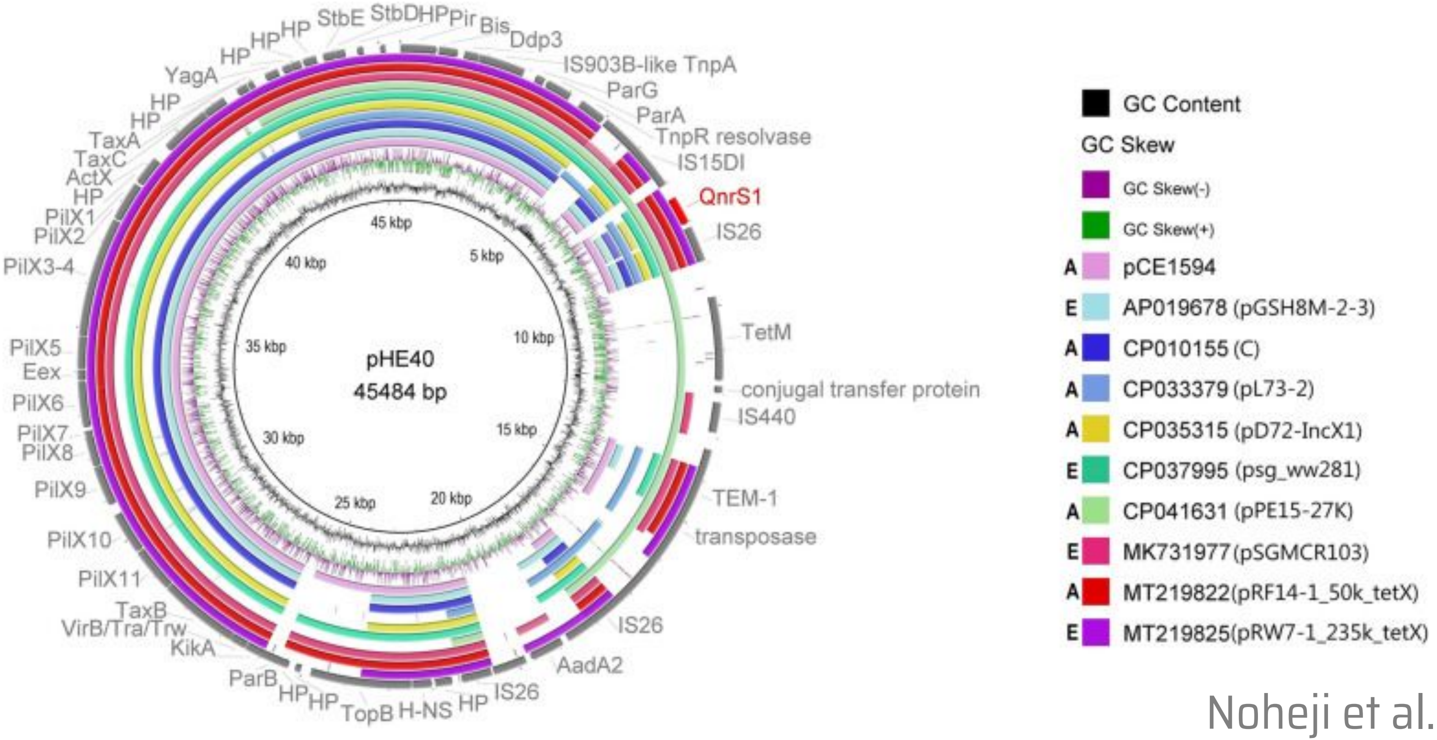
```
> chromosome
TAACCGATGTAACACTTTGCTCGTGCACATAGGCGTGACACCTAAGACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTTTTGA
CCGAGTCCGCTCCTTTCTAACTCACTGACTTCTCGCGGGCCATGTCCTGCACGCTCAACAGCCAGCGGTGTCCCGTTACCCTTCAAG
CTCATCTTCCTCGAGGTCTGTTGTAGTACCACACGCCTCTCCCGGCATATGCCCGCACTCCTCGACGGGACATTATGTGCCTTCAGTT
CCCGATCTCGGTCGCGCCCAGCCGGAATCCCTATGGCACCAGGGCCGCGTGAGCGAGAAGCGGGGGGAGAACTTTATAGGGCTGT
GGCTCATACAATAGGGTAAGGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCACCCACGGTATAAGACGAA
AGCCTAAGGCAACCCCGGGGTATGGCGCGGTCCTTCTCTCTACTTAACCGGTAAAACACTTTGCTCGTGCACATAGGCGTGACACCT
```

Annotated Genome



> chromosome
TAACCGA GCGTGACACCTAAGACTGGAACAAGCTCAGAAGAGTAGTAGGCGAGCATTTTTTGA
CCGAGTCCGCTCCTTTCTAACTCACTGACTTCTCGCGGGCCA
TACCACACGCCTCTCCCGGCATATGCCCGCACTCCTCGACGGGACATTATGTGCCTTCAGTT
CCCGATCTCGGTCGCGCCCAGCCGGAATCCC GCGAGAAGCGGGGGGAGAACTTTATAGGGCTGT
GGCTCATACAATAGGGTAAGGTTTCACCACATTTCTTCACTTCAGAAGCGACGCCTCCATTTTGCCCTCACCCACGGTATAAGACGAA
AGCCTAAGGCAACCCCGGGGT CCGGTAAAACACTTTGCTCGTGCACATAGGCGTGACACCT

What it may look like?



Noheji et al. 2022

Format of Annotation output

GFF3

```
                            start   end       strand         attributes
NC_002945.4 feature     gene  1524  2345  .     -     .     ID=gene:BQ2027_MB0001;Name=dnaA;biotype=protein_coding;
  seqid      source     type            score       phase
```

Format of Annotation output

GFF3

```
     CDS             complement(1524..2345)
                     /codon_start=1
                     /gene="BQ2027_MB0001"
                     /product="chromosomal replication initiator protein dnaA"
                     /label=dnaA
                     /note="activates initiation of DNA replication in bacteria."
                     /translation="MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGYI
                     ELDLNSGKILESFRPEERFPMMSTFKVLLCGAVLSRIDAGQEQLGRRIHYSQNDLVEYS
                     PVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRW
                     EPELNEAIPNDERDTTMPVAMATTLRKLLTGELLTLASRQQLIDWMEADKVAGPLLRSA
                     LPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGAS
                     LIKHW"
```

Format of Annotation output

GBK

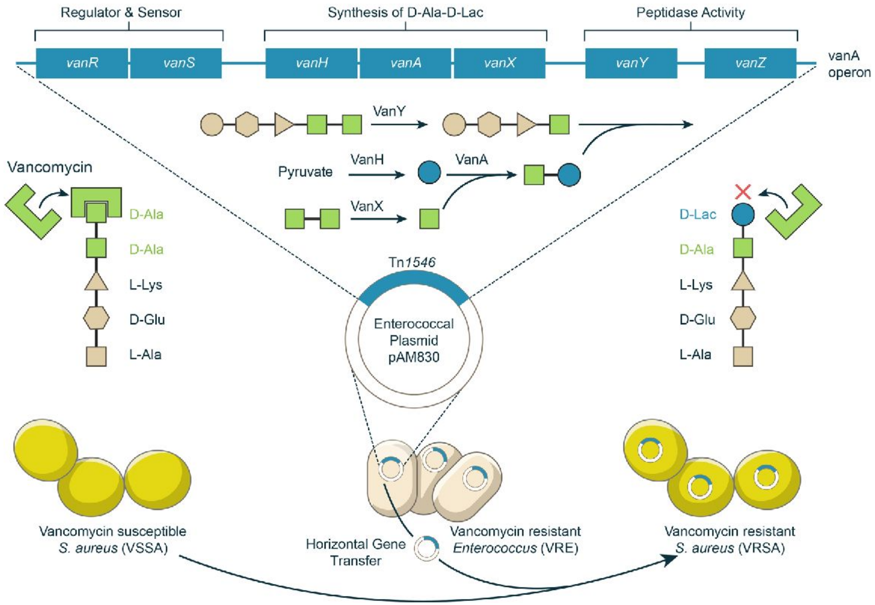
```
     CDS               complement(1524..2345)
                       /codon_start=1
                       /gene="BQ2027_MB0001"
                       /product="chromosomal replication initiator protein dnaA"
                       /label=dnaA
                       /note="activates initiation of DNA replication in bacteria."
                       /translation="MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGYI
                       ELDLNSGKILESFRPEERFPMMSTFKVLLCGAVLSRIDAGQEQLGRRIHYSQNDLVEYS
                       PVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRW
                       EPELNEAIPNDERDTTMPVAMATTLRKLLTGELLTLASRQQLIDWMEADKVAGPLLRSA
                       LPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGAS
                       LIKHW"
```

Challenges and Innovations

- Complexity of Microbial Genomes
  - Hard to assemble genomes
  - Repeat Regions
- Accuracy of Prediction Algorithms
  - Outdates / Incomplete databases
  - Atypical gene structures
  - Homology based errors
- New methods
  - Data Integration
  - Regular database updates
  - Unsupervised Machine Learning Models

Annotation Exercise

- Annotate three E. faecium genomes with Prokka and Proksee
- Explore annotation results
- Compare AMR potential



Thank you!

wellcome connecting science

ACORN