Preliminary Results

Team 1 Gene Prediction
Genevieve Brandt, Victoria Caban, Yuntian He, Junyu Li, Yiqiuyi Liu, Yihao Ou, Wenyi Qiu, Casey Smith, Mohit Thakur, Stephen Wist, Qinyu Yue

Content

Introduction

Reference-based

Ab-initio

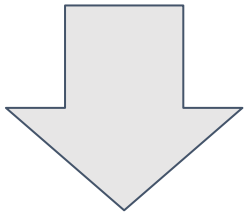
RNA prediction tools

Results and next steps

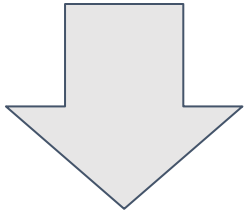
Content

Task

Use the assembled genomes from the previous group to predict genes and non-coding RNAs



Build a robust method that can be used on 262 genomes quickly and accurately

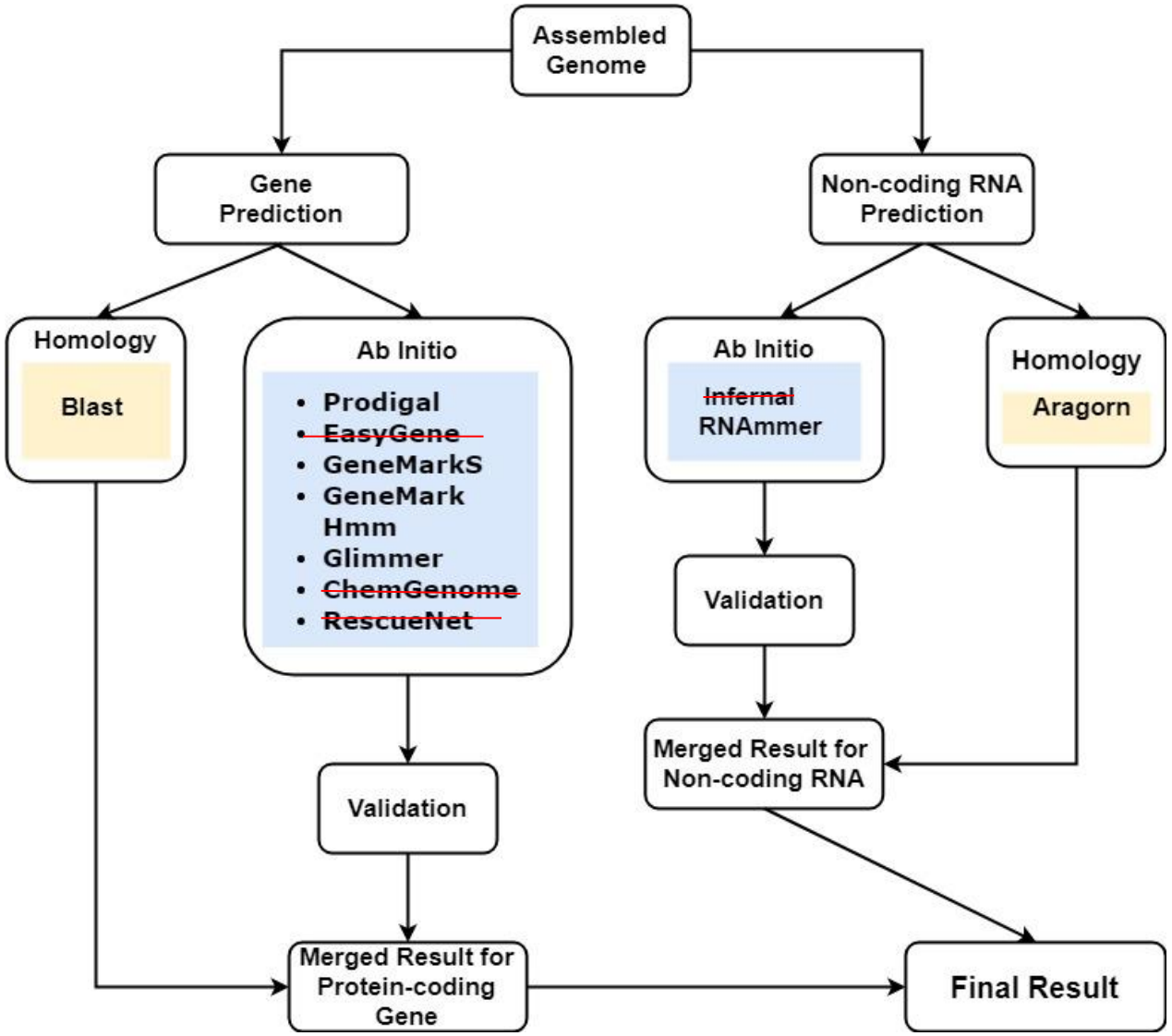


Understand the klebsiella genome and genomic elements and how they relate to heteroresistance

Overview of tools and pipeline

Tools	Algorithms
Prodigal	Dynamic programming gene finding
~~EasyGene~~	~~Hidden Markov Model~~
GeneMarkS	Hidden Markov Model
GeneMark HMM	Hidden Markov Model
~~Infernal~~	~~Hidden Markov Model~~
Glimmer	Interpolated Markov Model
RNAmmer	Markov Models
~~ChemGenome~~	~~Linear Discriminant Analysis~~
~~RescueNet~~	~~Synonymous codon usage~~
BLAST	BLAST
~~Aragorn~~	~~Heuristic tRNA detection~~

Overview of tools and pipeline



Content

Introduction

Reference-based

Ab-initio

RNA prediction tools

Results and next steps

Reference Selection

How did we select reference genomes?

- Not choosing a specific reference genomes
  - one genome against entire database
  - advantage: get all potential gene
  - disadvantage: slow

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Klebsiella pneumoniae subsp. pneumoniae KPNIH10, complete genome	4.152e+06	1.100e+07	99%	0.0	99%	CP007727.1
Klebsiella pneumoniae subsp. pneumoniae KPNIH1, complete genome	4.152e+06	1.098e+07	99%	0.0	99%	CP008827.1
Klebsiella pneumoniae strain AR_0113, complete genome	3.890e+06	1.050e+07	98%	0.0	99%	CP021751.1
Klebsiella pneumoniae isolate blood sample 2, complete genome	3.432e+06	1.097e+07	99%	0.0	99%	CP015822.1
Klebsiella pneumoniae strain AR_0112, complete genome	3.416e+06	1.075e+07	98%	0.0	99%	CP021549.1
Klebsiella pneumoniae isolate 207M1D0-sc-2013-04-03T11:21:06Z-1606409 genome assembly, chromosome: 1	3.080e+06	1.091e+07	97%	0.0	99%	LT216436.1

Reference Selection

How did we select reference genomes?

- The reference genomes used for assembling
  - one genome against one reference genome
  - advantage: accuracy
  - disadvantage: slow

qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_4014	100.000	4950	0	0	3986242	3991191	4950	1	0.0	9142
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_4890	100.000	4530	0	0	4837125	4841654	1	4530	0.0	8366
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1900	100.000	4449	0	0	1900144	1904592	1	4449	0.0	8216
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1870	100.000	4254	0	0	1861672	1865925	1	4254	0.0	7856
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1870	86.792	530	47	7	1862833	1863340	1090	1618	7.05e-159	569
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1870	86.792	530	47	7	1862761	1863289	1162	1669	7.05e-159	569
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1870	89.216	306	33	0	1862933	1863238	1118	1423	9.86e-103	383
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1870	89.216	306	33	0	1862789	1863094	1262	1567	9.86e-103	383
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1870	86.325	234	32	0	1863005	1863238	1118	1351	2.24e-64	255
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1870	86.325	234	32	0	1862789	1863022	1334	1567	2.24e-64	255
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_237	100.000	4224	0	0	231538	235761	1	4224	0.0	7801
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_3269	99.976	4104	1	0	3229533	3233636	4104	1	0.0	7574
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_5185	100.000	4053	0	0	5121243	5125295	4053	1	0.0	7485
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_236	100.000	4029	0	0	227433	231461	1	4029	0.0	7441
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1985	100.000	3953	0	0	1987588	1991540	3953	1	0.0	7300
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_2475	100.000	3903	0	0	2445753	2449655	1	3903	0.0	7208
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_4046	100.000	3888	0	0	4023136	4027023	3888	1	0.0	7180
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_1493	100.000	3882	0	0	1492632	1496513	1	3882	0.0	7169
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_4912	100.000	3798	0	0	4860036	4863833	3798	1	0.0	7014
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_503	100.000	3777	0	0	496796	500572	1	3777	0.0	6975
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_3263	100.000	3744	0	0	3217114	3220857	3744	1	0.0	6914
NZ_CP007727.1	lcl|NZ_CP007727.1_gene_3263	81.092	1227	209	21	2877504	2878720	1	1214	0.0	959

Reference: /projects/data/team1_genomeAssembly/reference_based_assembly/reference_genomes/GCF_000281435.2_ASM28143v2_genomic.fna
Query: blastn -db database -query /projects/data/team1_genomeAssembly/reference_based_assembly/assembly_50/SRR3982229/assembly.fasta

Reference Selection

How did we select reference genomes?

- Cluster 262 genomes into several groups
  - one group paired with a reference sequence
  - advantage: fast
  - disadvantage: not extremely accurate

Reference Selection

What commands and parameters did we use?

- local blast:
  - makeblastdb -in reference_genome.fa -dbtype nucl -out database_name
  - blastn -db database -query assembly.fasta -outfmt 6 -out result_name
- Mash
  - mash dist genome_a.fa genome_b.fa
- Parameter
  - BLAST: default
  - Mash Distance: undecided

Content

Introduction

Reference-based

Ab-initio

RNA prediction tools

Results and next steps

Ab-initio tools

- GeneMarkS
- GenMark.hmm
- Prodigal
- Glimmer

- command line
- parameters
- average time

GFF Gene Fiding Format

```
[qyue7@biogenome2018a GFF_format]$ head -20 SRR3982229.GFF
##gff-version 2
##source-version GeneMark.hmm_PROKARYOTIC 3.36
##date: Mon Feb 26 23:04:43 2018
# Sequence file name: ../../../../team1_genomeAssembly/reference_based_assembly/assembly_50/SRR3982
229/assembly.fasta
# Model file name: SRR3982229_hmm_combined.mod
# RBS: true
# Model information: GeneMarkS_gcode_11

NZ_CP007727.1   GeneMark.hmm    CDS     280     720     27.770355       -       0       gene_id=1
NZ_CP007727.1   GeneMark.hmm    CDS     820     1278    35.610959       -       0       gene_id=2
NZ_CP007727.1   GeneMark.hmm    CDS     1430    2422    63.394938       +       0       gene_id=3
NZ_CP007727.1   GeneMark.hmm    CDS     2426    3874    86.439334       -       0       gene_id=4
NZ_CP007727.1   GeneMark.hmm    CDS     3871    5370    94.220343       -       0       gene_id=5
NZ_CP007727.1   GeneMark.hmm    CDS     5588    7456    103.979517      +       0       gene_id=6
NZ_CP007727.1   GeneMark.hmm    CDS     7642    8061    18.724113       +       0       gene_id=7
NZ_CP007727.1   GeneMark.hmm    CDS     8072    9577    113.500344      +       0       gene_id=8
NZ_CP007727.1   GeneMark.hmm    CDS     9583    10548   57.990711       +       0       gene_id=9
NZ_CP007727.1   GeneMark.hmm    CDS     10576   11466   66.684727       +       0       gene_id=10
NZ_CP007727.1   GeneMark.hmm    CDS     11567   12499   55.101155       +       0       gene_id=11
NZ_CP007727.1   GeneMark.hmm    CDS     12512   13495   49.439073       +       0       gene_id=12
```

seqname source feature start end score strand frame attribute

GeneMarkS

- version: 4.32: April, 2015

- **`gmsn.pl --prok --name --output <output_name> --format GFF --fnn <input_file>`**

- default parameters:
  - --gcode 11: genetic code, 11 for the bacterial, archaeal and plant plastid code
  - --motif 1: true for iterative search for a sequence motif associated with CDS(coding DNA sequences) start
  - --prestart 40: <number> length of sequence upstream of translation initiation site that presumably includes the motif
  - --maxitr 10: maximum number of iterations
  - --identity 0.99: identity level assigned for termination of iterations

- output: **GFF**, GFF3, lst, **fasta.fnn** and fasta.faa etc.

- average time: 12 min / genome

GeneMarkS

```
#!/usr/bin/perl -w
use strict;
#define variables
my $filename = ();
my @SRRname = ();
#input file
$filename = @ARGV[0];
#check file if exit and open file
unless (-e $filename){
    print "This file \"$filename\" do not exit! Please check it!";
    exit;
}
unless (open FILENAME, $filename){
    print "Cannot open this file!!";
    exit;
}

@SRRname = <FILENAME>;
chomp @SRRname;
close FILENAME;
#run GeneMarkS, you can replace the command line in ``
foreach $i (@SRRname){
    `perl gmsn.pl --prok --output $i.GFF --format GFF --name $i ../../../../tea
    m1_genomeAssembly/reference_based_assembly/assembly_50/$i/assembly.fasta`
}
```

GeneMark.hmm

- version: 1.0: September, 2014
- **`perl gmhmmp.pl  --output <out put_name> <input_file>`**
- default parameters:
  - motif 1: true for iterative search for a sequence motif associated with CDS start
- output: GFF, lst, fasta.fnn, etc.
- average time: 10s / genome

Prodigal

- version: 2.6.3: February, 2016

- **`Prodigal -i [input_file] -o [output gene coordinates] -d [output nucleotide sequences] -a [output protein translations]`**

- default parameters:
  - translation table: standard bacteria/archaea table used first 11
  - gap-mode: partial genes can run into gaps
  - closed: not used (did not force closed end genes)
  - rbs-motif: default Shine-Delgarno used

- output: genbank (gbk), GFF format (gff), and simple coordinate output (SCO)

- average time: 17.011s / genome

Glimmer3

- version: 3.02 May 2006
- **`g3-from-scratch.csh <input_file> <output_file>`**
- Default parameters:
  - -o 50: max prediction overlap length
  - -g 110: min gene length
  - -t 30: max entropy distance score
  - -d 7 : depth of interpolated context model
  - -p 3: period of interpolated context model
  - -w 12: width of interpolated context model
- Output: gene tables and fasta
- Average time: 53.892s / genome

Ab-initio

overview of ab-initio tools

	Prodigal	Glimmer3	GeneMarkS	GeneMark.hmm
average time	17.011s	53.892s	12.11 min	10s
output	GFF and fata	fasta (GFF can be converted from the gene table)	GFF and fasta.fnn	GFF and fasta.fnn
parameters	-translation table 11 -rbs-motif: Shine-Delgarno -gap-mode -closed	-o 50: max prediction overlap length -g 110: min gene length -t 30: max entropy distance score -p 3: period of interpolated context model	-gcode 11 -motif 1 -prestart 40 -maxitr 10 -identity 0.99	-motif 1

Content

Introduction

Reference-based

Ab-initio

RNA prediction tools

Results and next steps

RNA tools

- Infernal
- Aragorn
- RNAmmer

RNA tools

- ~~Infernal~~
- ~~Aragorn~~
- RNAmmer

RNAmmer

- Command:

```
./rnammer -s bac -m lsu,ssu,tsu -multi -gff output.gff -f output.fasta
-h output_report.html < input.fasta
```

- -s: kingdom

  -m: Molecule type

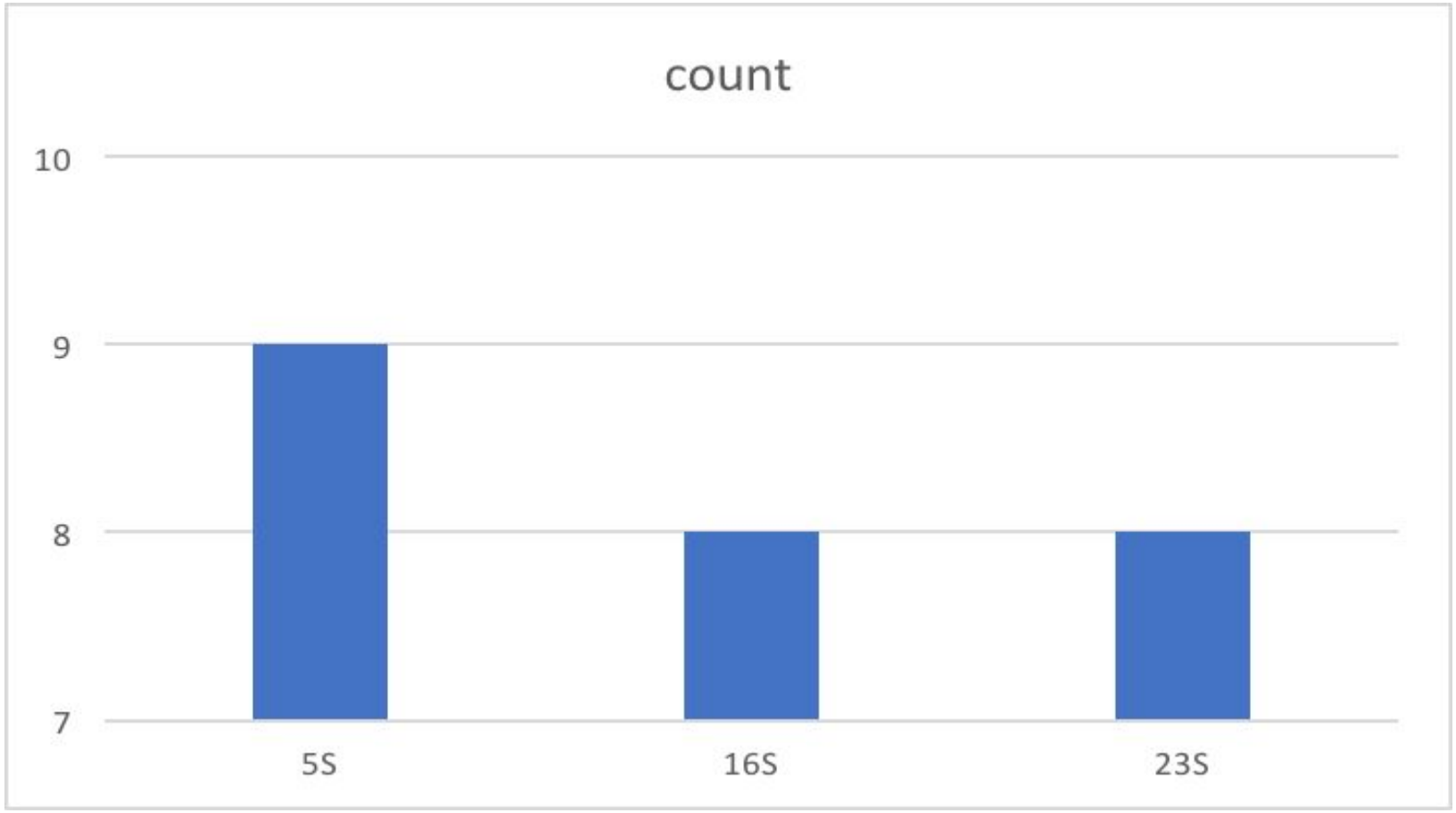
  -multi: Runs all molecules and both strands in parallel

- Time: ~48s

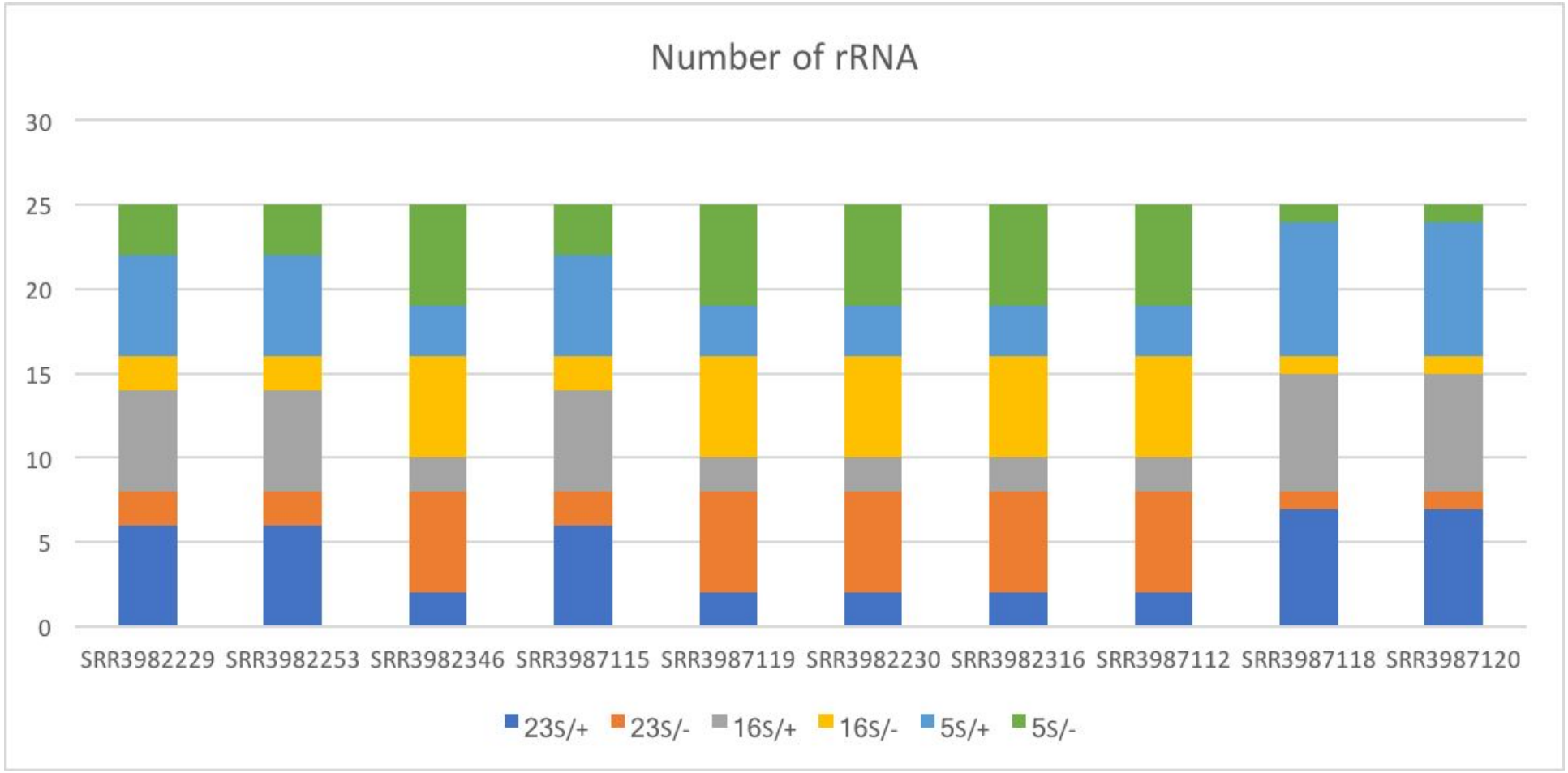
RNA tools

```
##gff-version2
##source-version RNAmmer-1.2
##date 2018-02-27
##Type DNA
# seqname           source                        feature     start       end    score   +/-  frame  attribute
# ---------------------------------------------------------------------------------------------------------
NZ_CP008797.1    RNAmmer-1.2     rRNA    1005793 1008692 3728.2  +       .       23s_rRNA
NZ_CP008797.1    RNAmmer-1.2     rRNA    214540  217439  3729.1  +       .       23s_rRNA
NZ_CP008797.1    RNAmmer-1.2     rRNA    1797527 1800426 3733.5  +       .       23s_rRNA
NZ_CP008797.1    RNAmmer-1.2     rRNA    122496  125395  3730.0  +       .       23s_rRNA
NZ_CP008797.1    RNAmmer-1.2     rRNA    4904412 4907311 3727.8  -       .       23s_rRNA
NZ_CP008797.1    RNAmmer-1.2     rRNA    259589  262488  3727.5  +       .       23s_rRNA
NZ_CP008797.1    RNAmmer-1.2     rRNA    17965   20864   3726.1  +       .       23s_rRNA
```

RNA tools



RNA tools


Number of rRNA
30
25
20
15
10
5
0
SRR3982229
SRR3982253
SRR3982346
SRR3987115
SRR3987119
SRR3982230
SRR3982316
SRR3987112
SRR3987118
SRR3987120
23S/+
23S/-
16S/+
16S/-
5S/+
5S/-

RNA tools -- Next Step

- Test on tRNAscan-SE 2.0 (December 2017)
  - Identify ~ 99% true tRNA
  - < 1 false positive per 15 billion nucleotide

S. Jaenicke

Content

Introduction

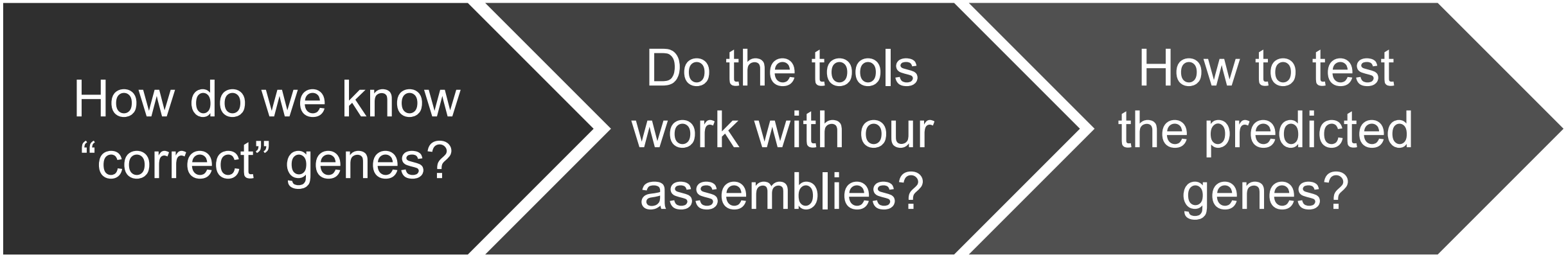
Reference-based

Ab-initio

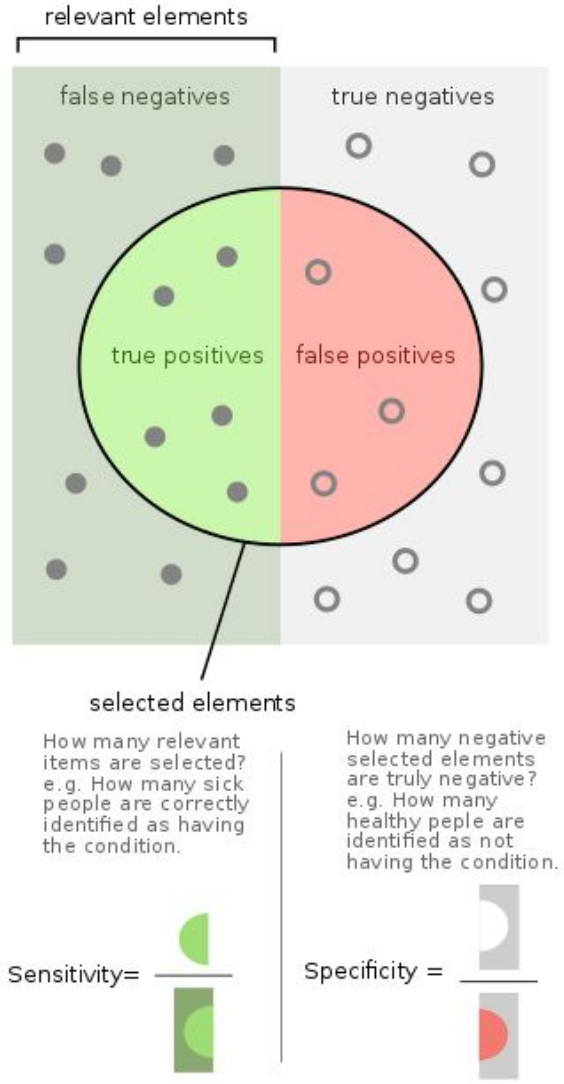
RNA prediction tools

Results and next steps

Choosing tools for gene prediction



Sensitivity vs Specificity



Sensitivity

- True positive rate
- Probability of detection
- What % of genes are correctly identified as genes?

Specificity

- True negative rate
- What % of “not genes” are correctly identified as “not genes”?

Our measures of accuracy

Sensitivity

- True positives over true positives plus false negatives
- TP / (TP + FN)

Specificity

- True negatives over true negatives plus false positives
- TN / (TN + FP)
- This is not possible for us to calculate, because we would need to have a value for things identified as “not genes” which does not apply to what we are analyzing
- We do not know the true negative value in the above equation

Positive predictive value (PPV)

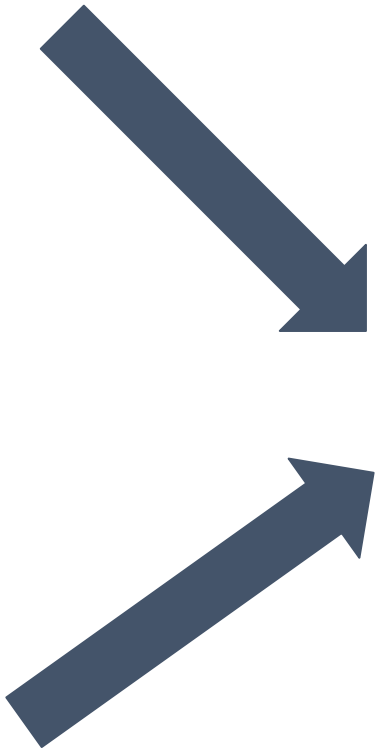
- Is a positive result actually a positive?
- PPV = TP/ (TP + FP)

Our measures of accuracy

10 assembled genomes from 1st group

+

Reference genes from NCBI based on the MASH tree



BLAST predicted genes against reference

True positives: number of predicted genes that match reference
False positives: number of predicted genes that do not match the reference

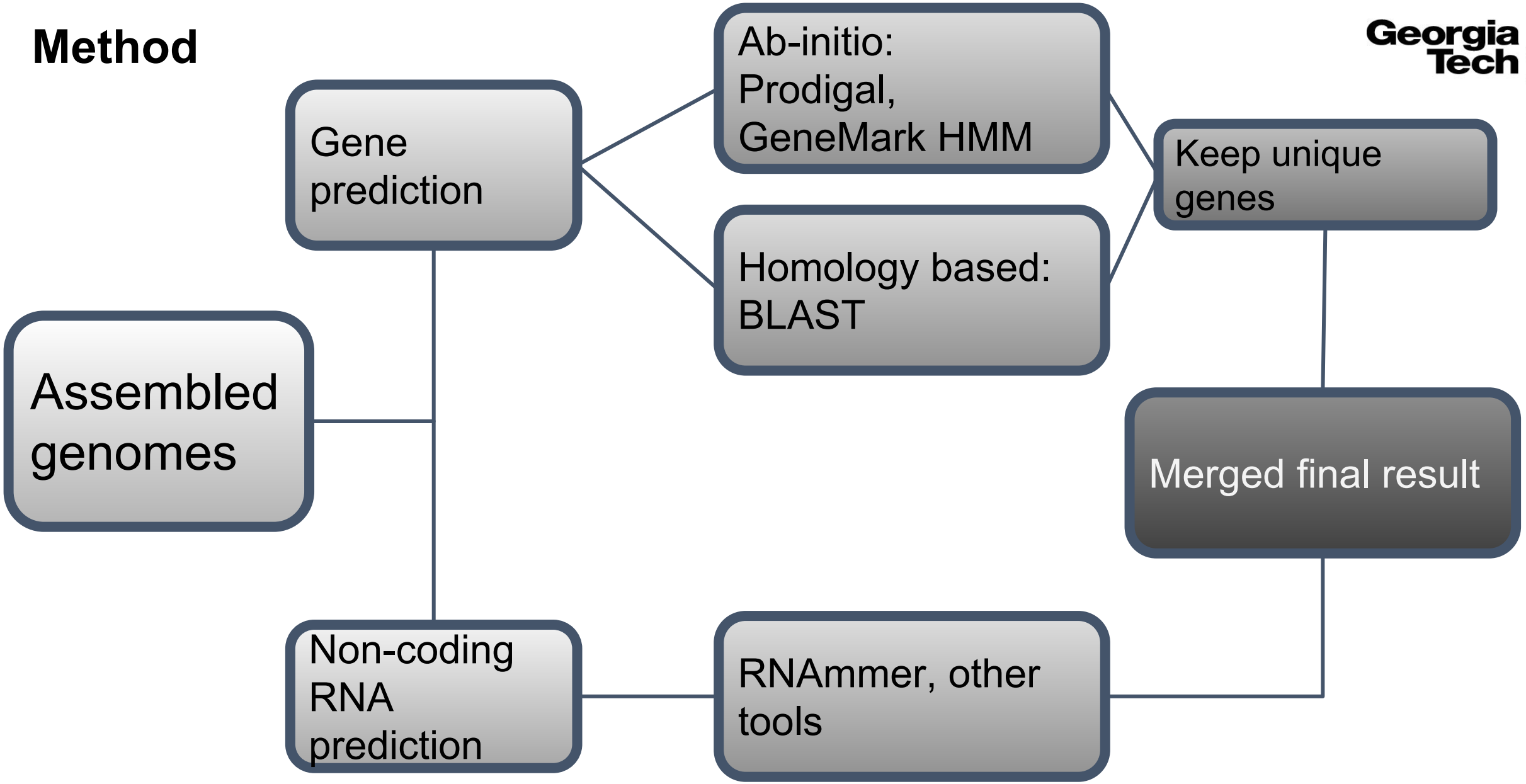
BLAST reference against predicted genes

False negatives: number of reference genes not in the prediction

Comparing our tools

	Sensitivity (TP / (TP + FN))	PPV (TP / (TP + FP))	Run time
Glimmer	93.47	96.36	~54 seconds
GeneMark S	93.10	91.24	~12 min
GeneMark HMM	93.11	93.10	~10 seconds
Prodigal	94.71	94.07	17 seconds

Method


Gene prediction
Ab-initio: Prodigal, GeneMark HMM
Homology based: BLAST
Keep unique genes
Assembled genomes
Merged final result
Non-coding RNA prediction
RNAmmer, other tools

Questions?

(The homework is now posted on the wiki)