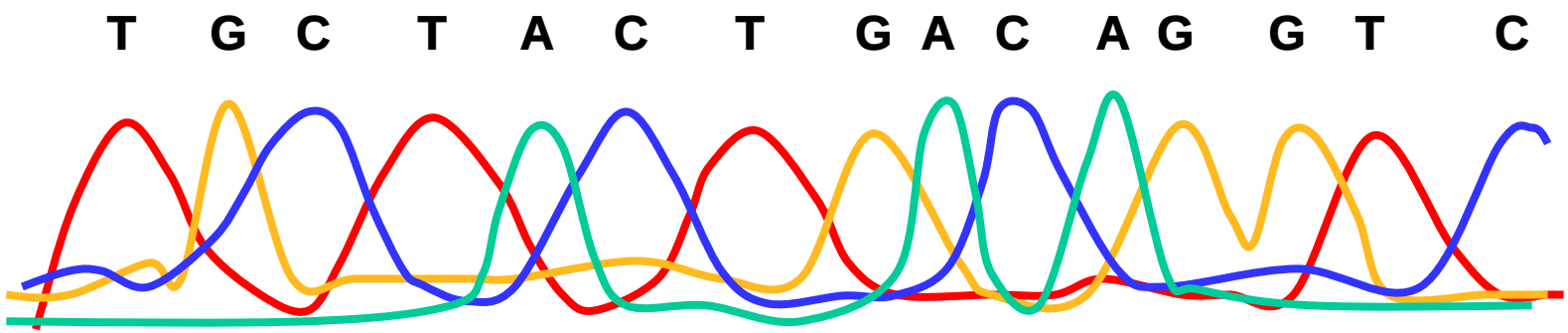
Fine-scale sequence polymorphisms and genome assembly

Steven L. Salzberg*

Pathema Bioinformatics Resource Center, TIGR

*As of July 1: Center for Bioinformatics and Computational Biology, University of Maryland

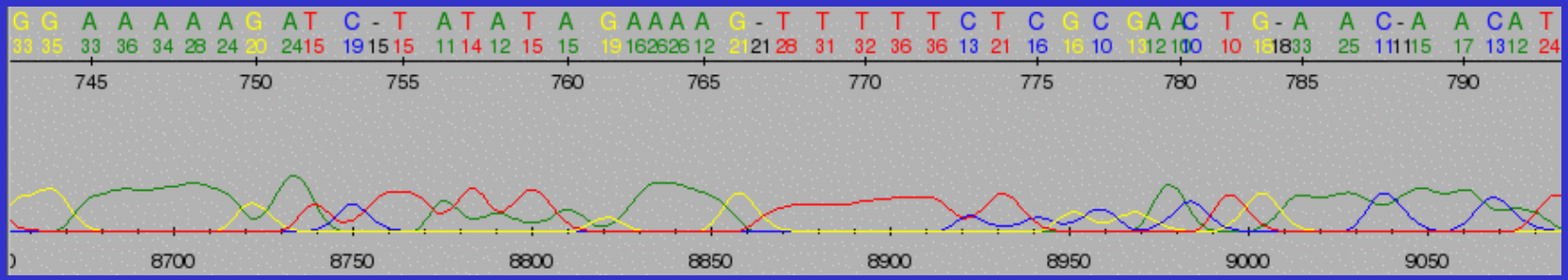
How accurate is sequencing?



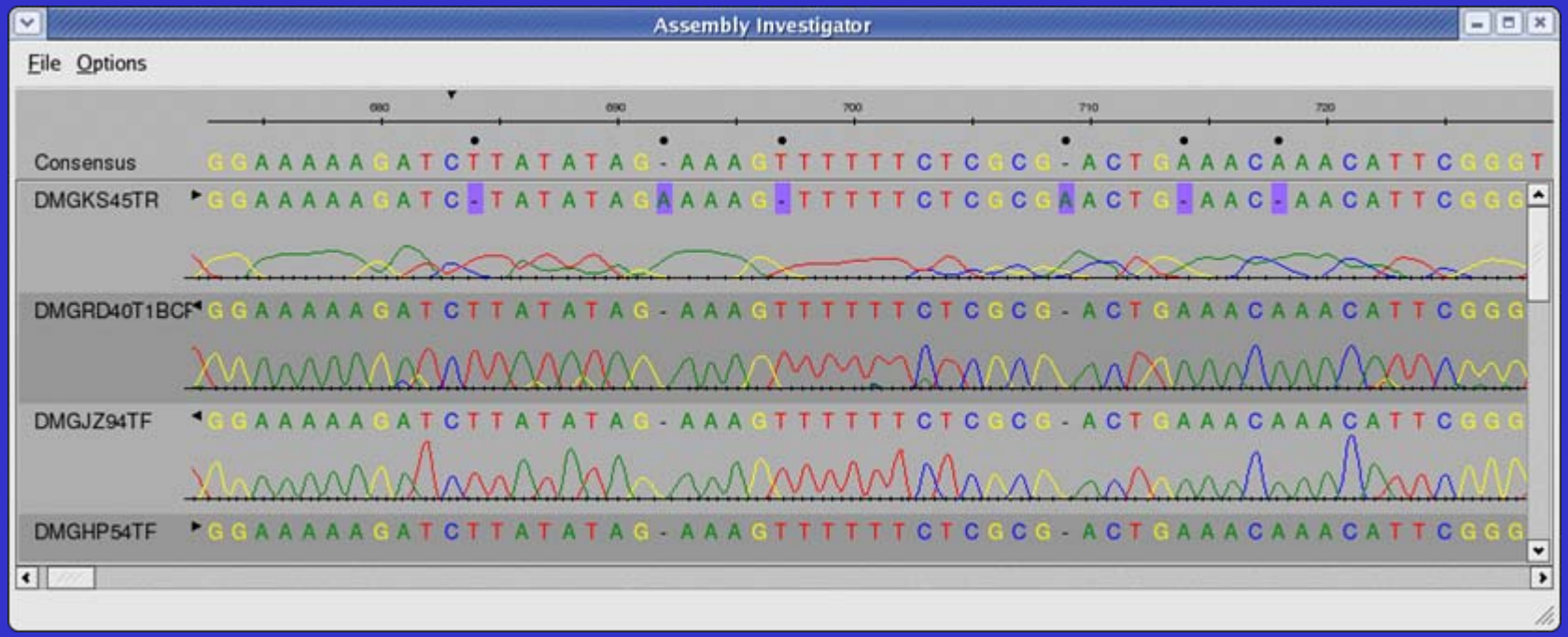
99%, 800 bases

AutoEditor: a 2nd-generation basecaller

Base-calling in the context of single chromatogram is hard…



but finding base-calling “mistakes” in a multiple alignment is easy.



Nucleic Acids Research

HOME | HELP | FEEDBACK | SUBSCRIPTIONS | ARCHIVE | SEARCH | TABLE OF CONTENTS

QUICK SEARCH: [advanced]
Author: Keyword(s):
Go
Year: Vol: Page:

OXFORD Journals online

Published online 26 January 2004
Nucleic Acids Research, 2004, Vol. 32, No. 2 **562-569**
© 2004 Oxford University Press

- Full Text of this Article
- Reprint (PDF) Version of this Article
- Email this article to a friend
- Similar articles found in:
  Nucl. Acids. Res. Online
  PubMed
- PubMed Citation
- Search PubMed for articles by:
  Gajer, P. || Salzberg, S. L.
- Alert me when:
  new articles cite this article
- Download to Citation Manager

Automated correction of genome sequence errors

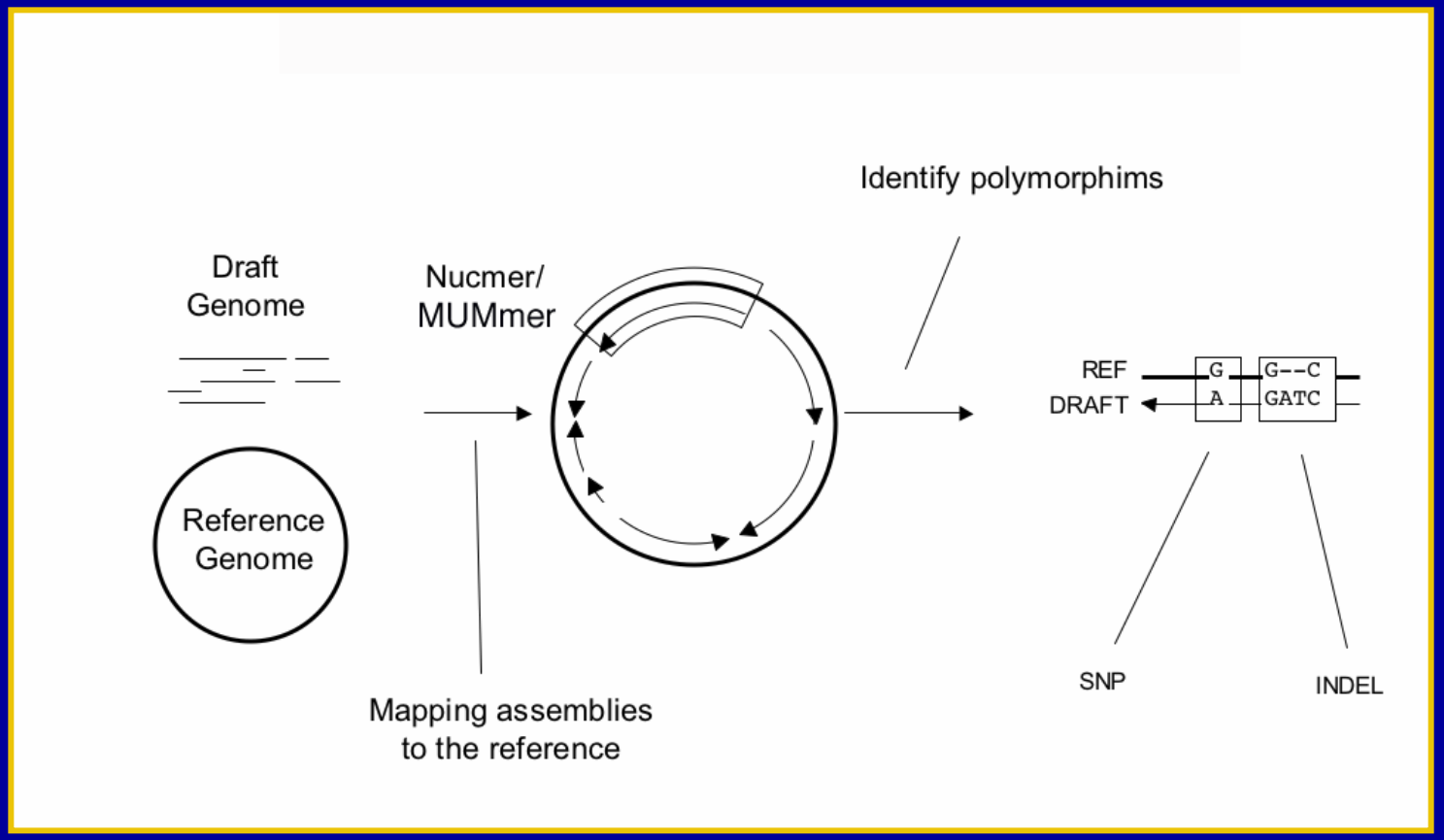
Pawel Gajer*, Michael Schatz and Steven L. Salzberg

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

*To whom correspondence should be addressed. Tel: +1 301 795 7854; Fax: +1 301 795 7208; Email: pgajer@tigr.org

By using information from an assembly of a genome, a new program called AutoEditor significantly improves base calling accuracy over that achieved by previous algorithms. This in turn improves the overall accuracy of genome sequences and facilitates the use of these sequences for polymorphism discovery. We describe the algorithm and its application in a large set of recent genome sequencing projects. The number of erroneous base calls in these projects was reduced by 80%. In an analysis of over one million corrections, we found that AutoEditor made just one error per 8828 corrections. By substantially increasing the accuracy of base calling, AutoEditor can dramatically accelerate the process of finishing genomes, which involves closing all gaps and ensuring minimum quality standards for the final sequence. It also greatly improves our ability to discover single nucleotide polymorphisms (SNPs) between closely related strains and isolates of the same species.

Polymorphism Discovery Pipeline

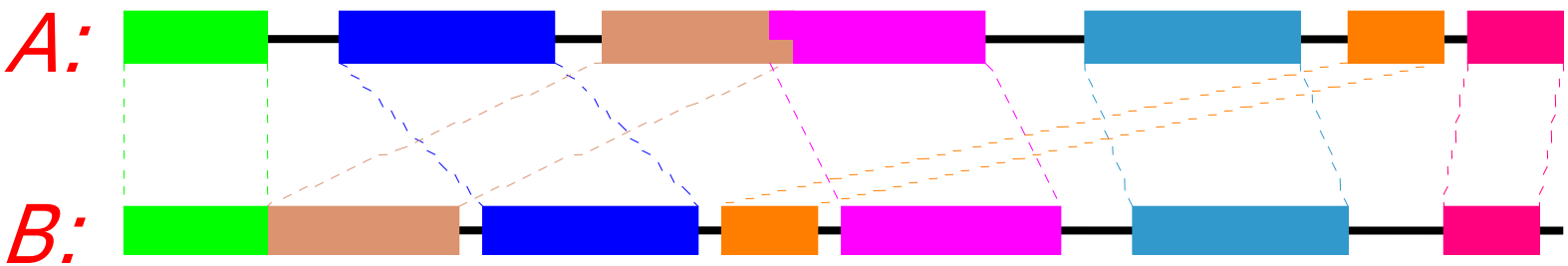


Quality Assessement of Polymorphism:

High Quality Polymorphism:
Coverage ≥ 3X and Sequence Quality Value ≥ 30 on both genomes

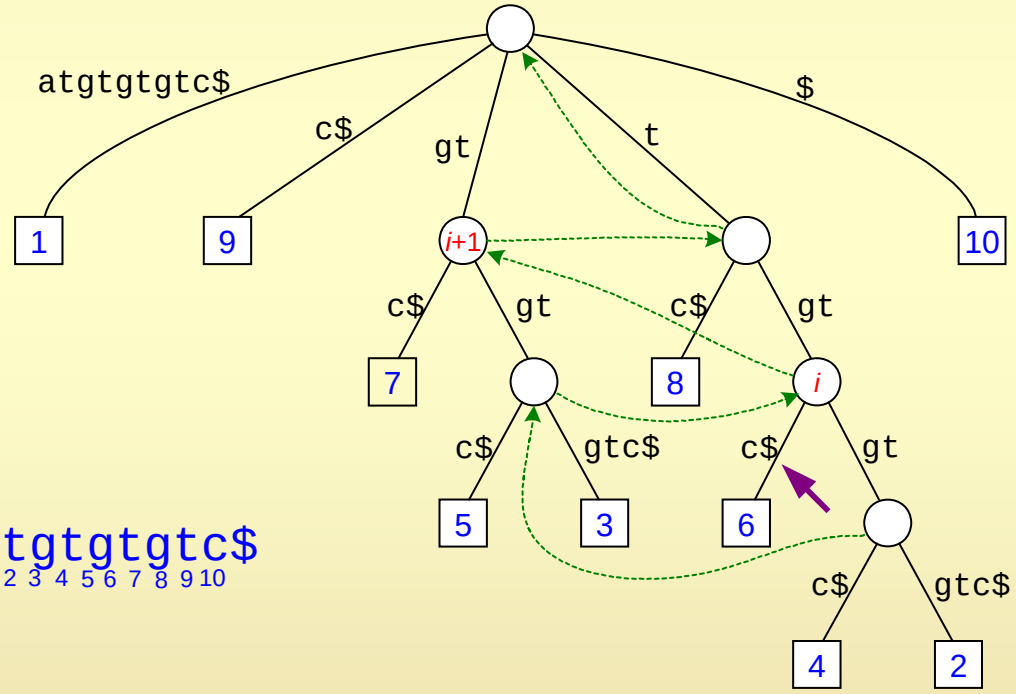
High-speed alignment of whole genomes

- Solution: MUMmer (Delcher et al., *Nucleic Acids Res.*, 2002) http://www.tigr.org/software/mummer
- Alignment of two 5 Mb genomes: 11 sec
- Aligning hundreds of contigs from incomplete genomes: < 1 minute
- Linear time **AND** space requirements



MUMmer's suffix tree algorithm

Streaming String **...atgtcc...**



Suffix Tree for String atgtgtgtc$
1 2 3 4 5 6 7 8 9 10

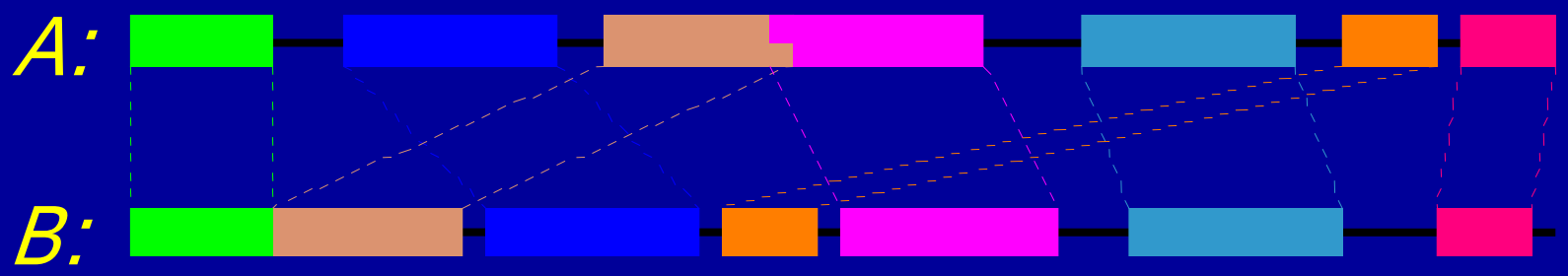
© 2004 Steven L. Salzberg

MUMmer 1.0, 2.0, 3.0

MUMmer 1: Sequences in genomes *A* and *B* that:

Occur *exactly once* in *A* and in *B*

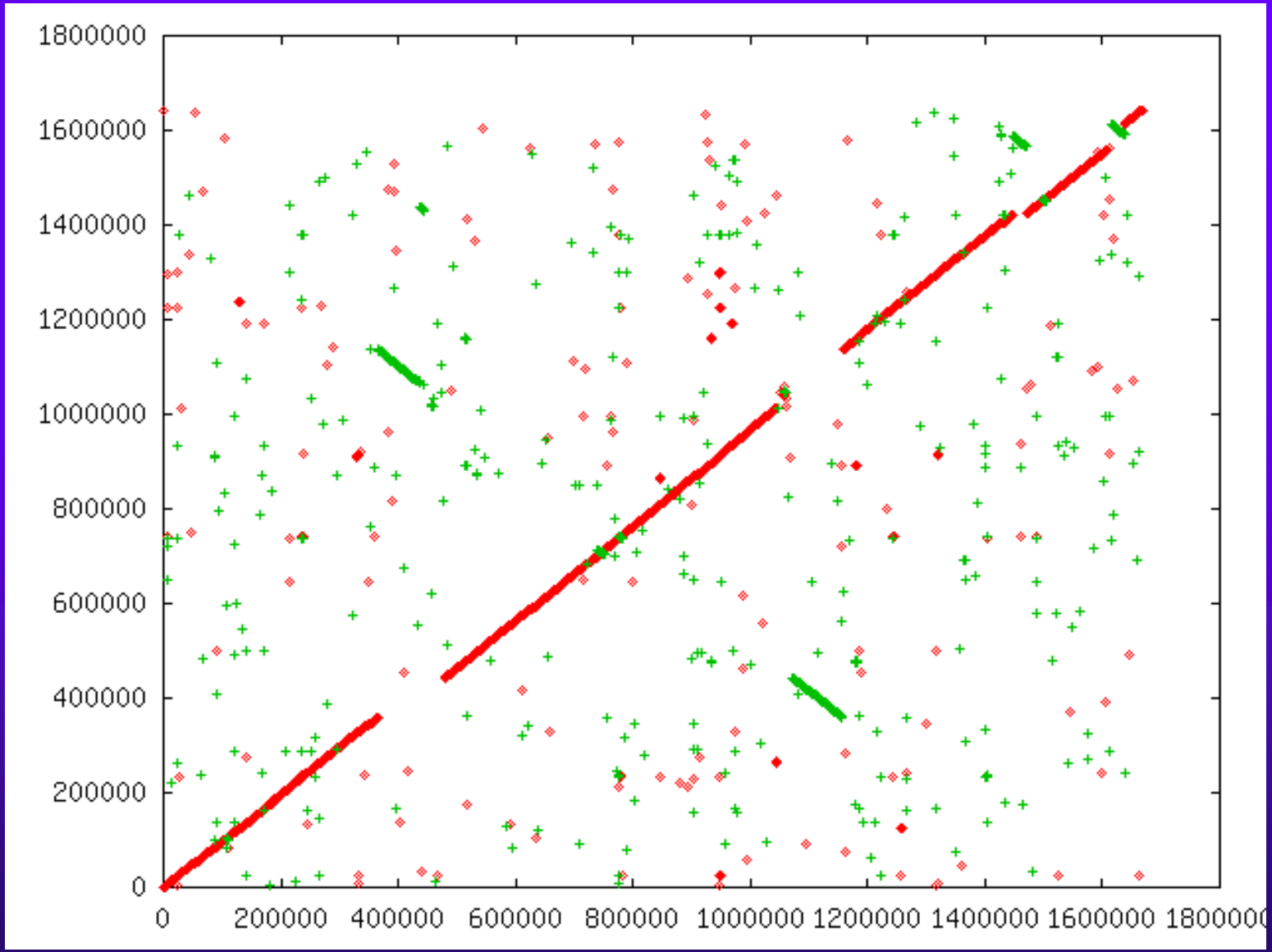
Are not contained in any larger matching sequence



MUMmer 2 and 3:

- Sequences not necessarily unique in A or B
- Sequences can be protein, DNA, English text

H. pylori 26695 vs. J99



http://www.tigr.org/software/mummer/manual/

MUMMER 3 MANUAL

Table of Contents

1. Introduction
    1. Description
    2. Comparative genomics
        1. Available sequence
        2. Human vs. Human
    3. OSI open source
2. Installation
    1. System requirements
    2. Obtaining MUMmer
    3. Compilation and installation
3. Running MUMmer
4. Use cases and walk-throughs
    1. Aligning two finished sequences
        1. Highly similar sequences without rearrangements
        2. Highly similar sequences with rearrangements
        3. Fairly similar sequences
        4. Fairly dissimilar sequences
    2. Aligning two draft sequences
    3. Mapping a draft sequence to a finished sequence
    4. Identifying repeats
5. Program descriptions
    1. Maximal exact matching
        1. mummer
        2. repeat-match
        3. exact-tandems
    2. Clustering
        1. gaps
        2. mgaps
    3. Alignment generators
        1. NUCmer
        2. PROmer

After correcting errors, find SNPs....

- Align genomes with MUMmer
- Find all differences:
  - SNPs
  - Insertions and deletions
  - Rearrangements
- Link SNPs and indels to assemblies

SNP quality computation

B. anthracis Ames Florida

consensus	TGAATGCACAC	T	G	A	A	T	**G**	C	A	C	A	C
Covering reads:												
GBZEI27TF	TGAATGCACAC	26	30	34	36	33	36	36	37	36	36	36
GBXEZ08TR	TGAATGCACAC	27	30	33	35	41	37	36	23	36	36	36
GBZDA09TF	TGAATGCACAC	26	18	35	31	26	20	29	19	36	36	36
cumulative quality		79	60	102	101	100	92	99	60	108	108	108
coverage		3	2	3	3	3	3	3	2	3	3	3

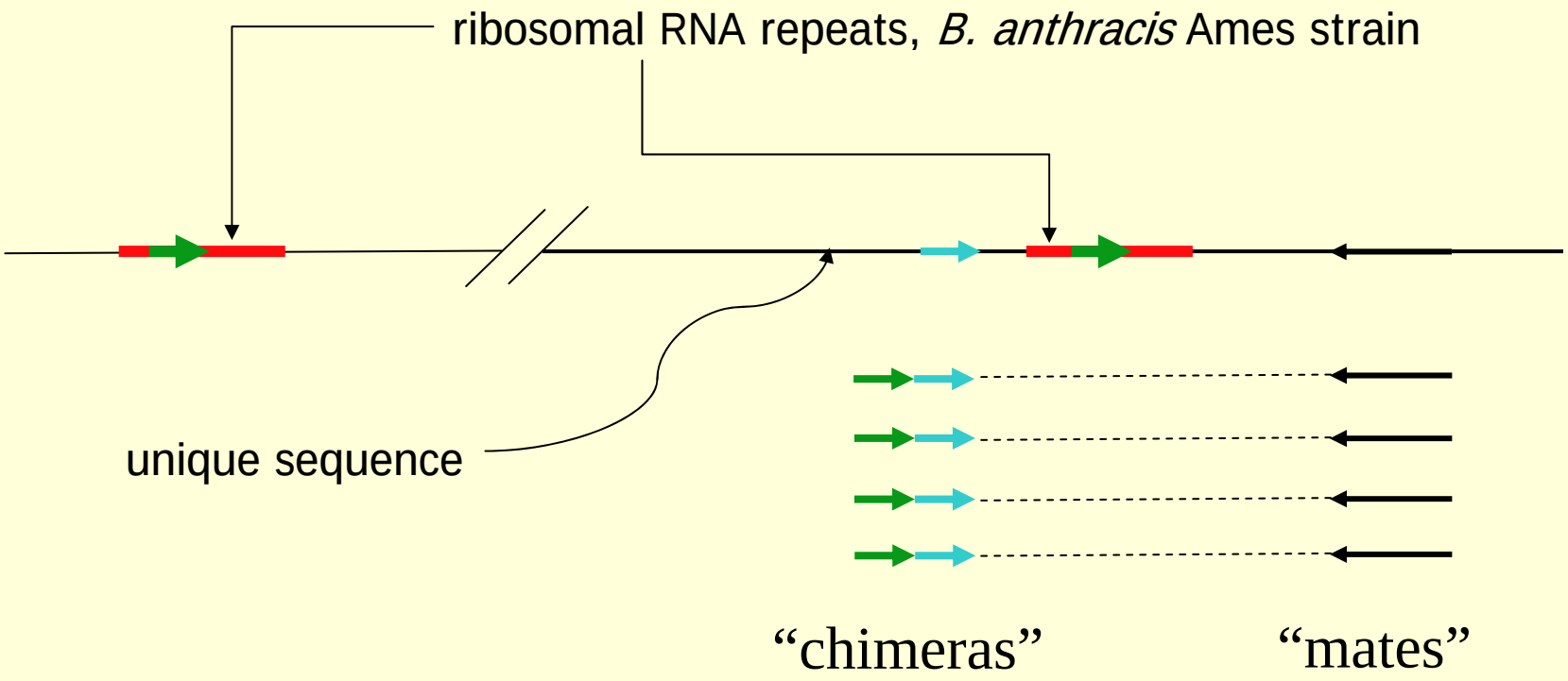
B. anthracis Ames Porton

consensus	TGAATACACAC	T	G	A	A	T	**A**	C	A	C	A	C
Covering reads:												
GBIFW80TF	TGAATA-ACAC	34	33	15	16	13	09	00	11	11	11	15
GBICA33TR	TGAATACACAC	34	35	34	35	35	34	34	36	36	36	36
GBIFQ32TR	TGAATACACAC	10	13	12	18	24	13	21	12	13	14	13
GBICH40TF	TGAATACACAC	36	36	36	32	32	32	32	32	31	31	36
GBICU19TR	TGAATACACAC	21	30	33	18	19	24	21	11	36	10	29
cumulative quality		125	134	103	67	91	90	108	68	103	67	101
coverage		4	4	4	2	3	3	4	2	3	2	3

Quality value = -10 $\log_2$ (probability of error)

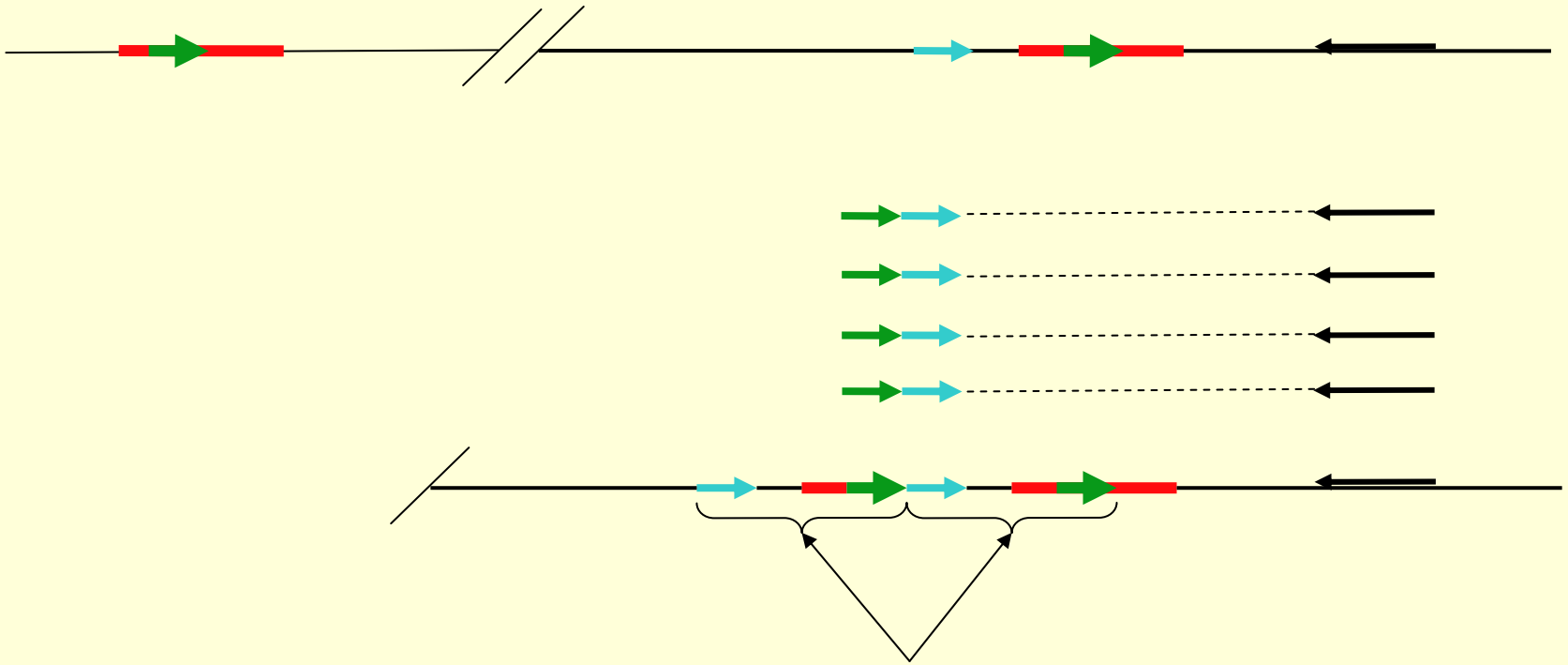
qv = 20 corresponds to 1 error in 100 bp

The importance of capturing traces: fixing mis-assemblies (even after publication)



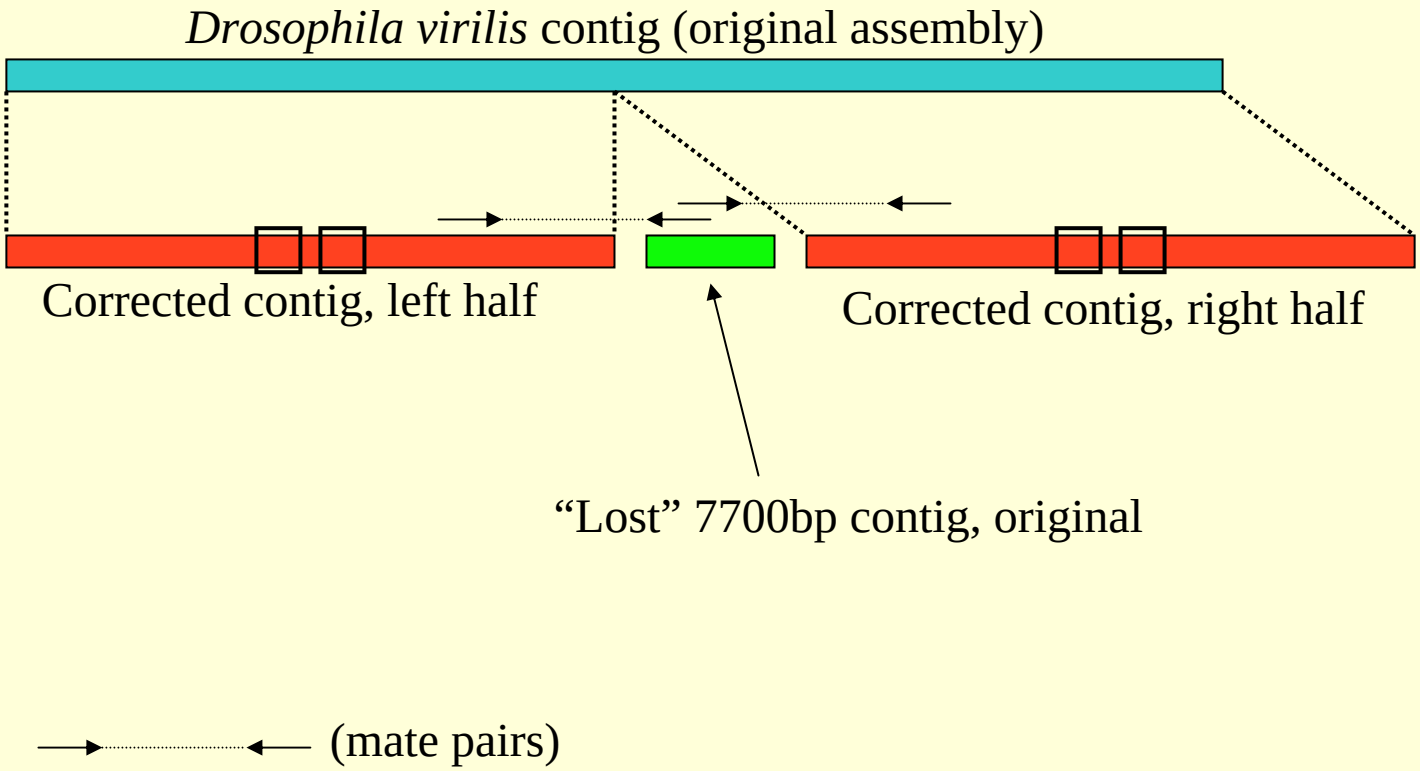
How do you fit the “orphan” green/blue sequences into the assembly?

Fixing the *B. anthracis* mis-assembly



Newly discovered tandem duplication

The importance of capturing traces II: Finding and correcting mis-assemblies



BRC goals include:

1. Capture assemblies of all pathogens (if possible)
2. Deposit assemblies in Assembly Archive

http://www.ncbi.nih.gov/Traces/assembly/assmbrowser.cgi

Community Page

The Genome Assembly Archive: A New Public Resource

Steven L. Salzberg, Deanna Church, Michael DiCuccio, Eugene Yaschenko, James Ostell

Scientists have dedicated considerable effort to decoding the genomes of an ever-growing list of species, ranging from small viruses, whose genomes may be just a few thousand nucleotides in length, to large mammalian genomes, three billion nucleotides and larger. Many aspects of life science research have benefited by the accumulation of these data, but decoded genomes could be even more valuable if important information about the genome sequence, currently being

a time; genomes must therefore be shattered into many small fragments (in what is known as the "shotgun" approach), which are then sequenced. The sequences are assembled to generate a consensus sequence that, if all steps work perfectly, matches the original DNA molecule. Since the sequencing of *Haemophilus influenzae* in 1995 (Fleischmann et al. 1995), most bacterial and archaeal species have been sequenced by fragmenting the entire genome, sequencing the pieces, and assembling the result (the whole-

the end product: the final genome sequence itself. It is only natural to regard the genome sequence, which is the basis for gene discovery and for functional understanding of the biology of the organism, as the primary result of a WGS project. In reality, though, a WGS project is an experiment in which large numbers of sequencing reactions are run, followed by a combination of computational work and additional

Citation: Salzberg SL, Church D, DiCuccio M, Yaschenko E, Ostell J (2004) The genome assembly archive: A new public resource. PLoS Biol 2(9): e285.

PLoS BIOLOGY
a peer-reviewed, open-access journal
PUBLIC LIBRARY of SCIENCE

Assembly Archive progress

Assembly (genome) name	Submitter	Length (bp)	# Contigs	# Traces	Coverage
Bacillus anthracis Ames Ancestor	TIGR	5,503,926	3	98,699	12.6X
Bacillus anthracis str. A1055	TIGR	5,370,618	42	83,857	12.9X
Bacillus anthracis str. Australia 94	TIGR	5,502,894	49	94,048	14.5X
Bacillus anthracis str. CNEVA-9066	TIGR	5,486,927	30	88,891	13.0X
Bacillus anthracis str. Kruger B	TIGR	5,470,007	64	89,421	11.8X
Bacillus anthracis str. Vollum	TIGR	5,487,718	52	102,108	15.7X
Bacillus anthracis str. WNA USA6153	TIGR	5,510,369	44	91,115	12.3X
Brucella suis 1330	TIGR	3,315,175	2	37,025	7.9X
Burkholderia mallei ATCC 23344	TIGR	5,835,527	2	82,118	8.2X
Campylobacter coli RM2228	TIGR	1,860,666	38	20,798	8.1X
Campylobacter jejuni RM1221	TIGR	1,777,831	1	24,334	8.4X
Campylobacter lari RM2100	TIGR	1,562,926	7	31,947	16.3X
Campylobacter upsaliensis RM3195	TIGR	1,770,638	19	22,615	9.1X
Coxiella burnetii RSA 493	TIGR	2,032,654	2	39,699	14.1X
Staphylococcus aureus pLW043	TIGR	57,889	1	958	12.8X
Staphylococcus epidermidis RP62A	TIGR	2,643,840	2	58,537	14.9X
Streptococcus agalactiae 2603V/R	TIGR	2,160,267	1	31,959	10.4X
Streptococcus agalactiae A909	TIGR	2,127,839	1	30,129	9.0X
Trichomonas vaginalis	TIGR	178,334,557	9,717	1,050,648	5.5X

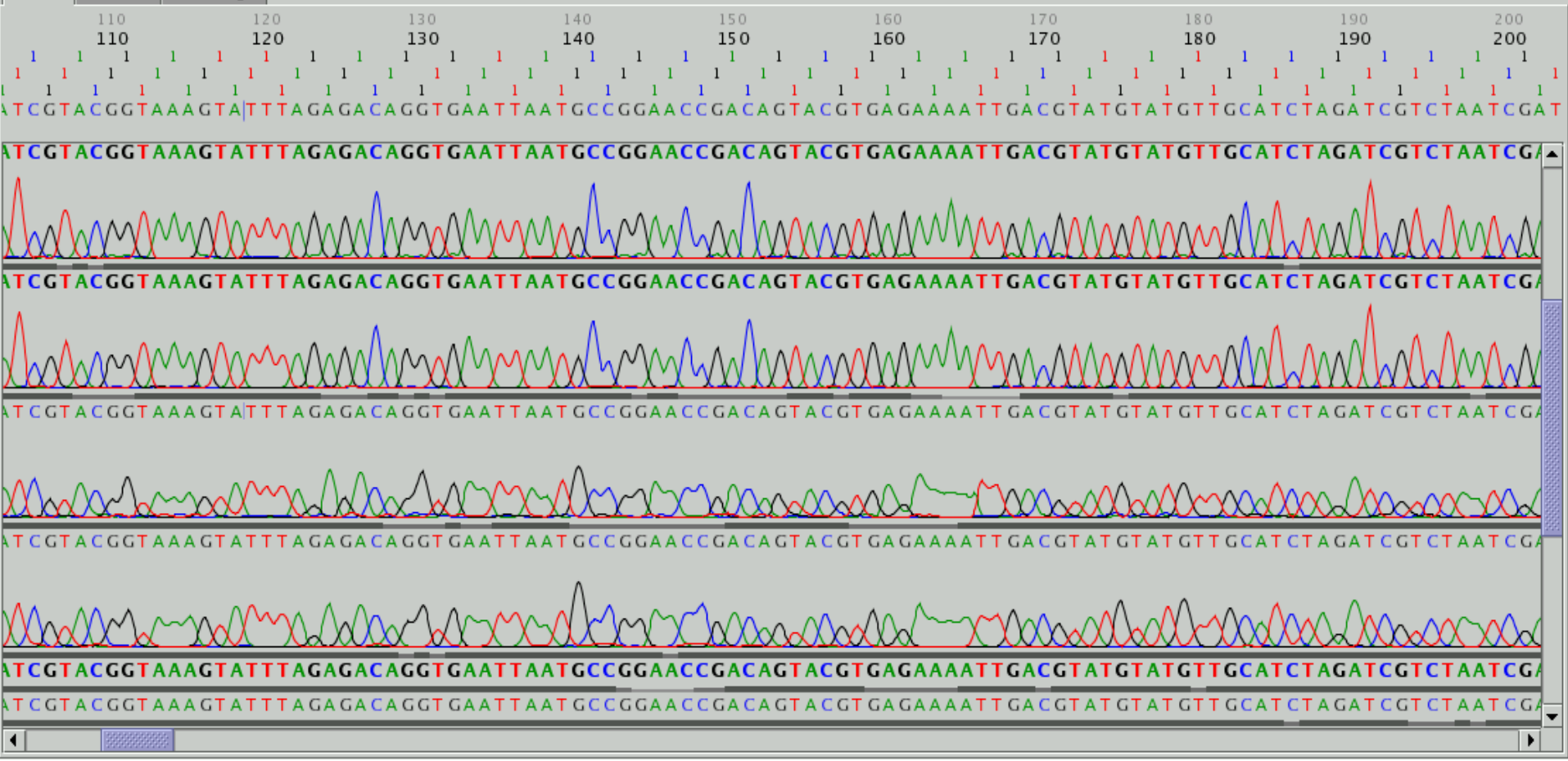
File Edit View Windows QA

Assembly Sequence List

Excerpt from a 12X anthrax assembly region...

Dirty?								fy date
☐	BAE							30 14:29:36...
☐	BAF							30 14:29:36...
☐	BAE							30 14:29:36...
☐	BAEH936TR	☑	-294	473	768	2003-06-24 09:32:32...	autoEdit	2003-12-30 14:29:36...
☐	BAFIH40TF	☑	-280	520	801	2002-11-05 15:06:13...	autoEdit	2003-12-30 14:29:36...
☐	BAEDA49TF	☐	-266	345	612	2002-10-08 21:36:27...	autoEdit	2003-12-30 14:29:36...
☐	BAFFG77TF	☐	-247	267	515	2002-10-15 11:07:59...	autoEdit	2003-12-30 14:29:36...
☐	BAFHU73TF	☑	-243	406	650	2002-11-05 13:08:26...	autoEdit	2003-12-30 14:29:36...
☐	BAEBV21TRB	☑	-208	587	796	2002-10-10 07:35:35...	autoEdit	2003-12-30 14:29:36...

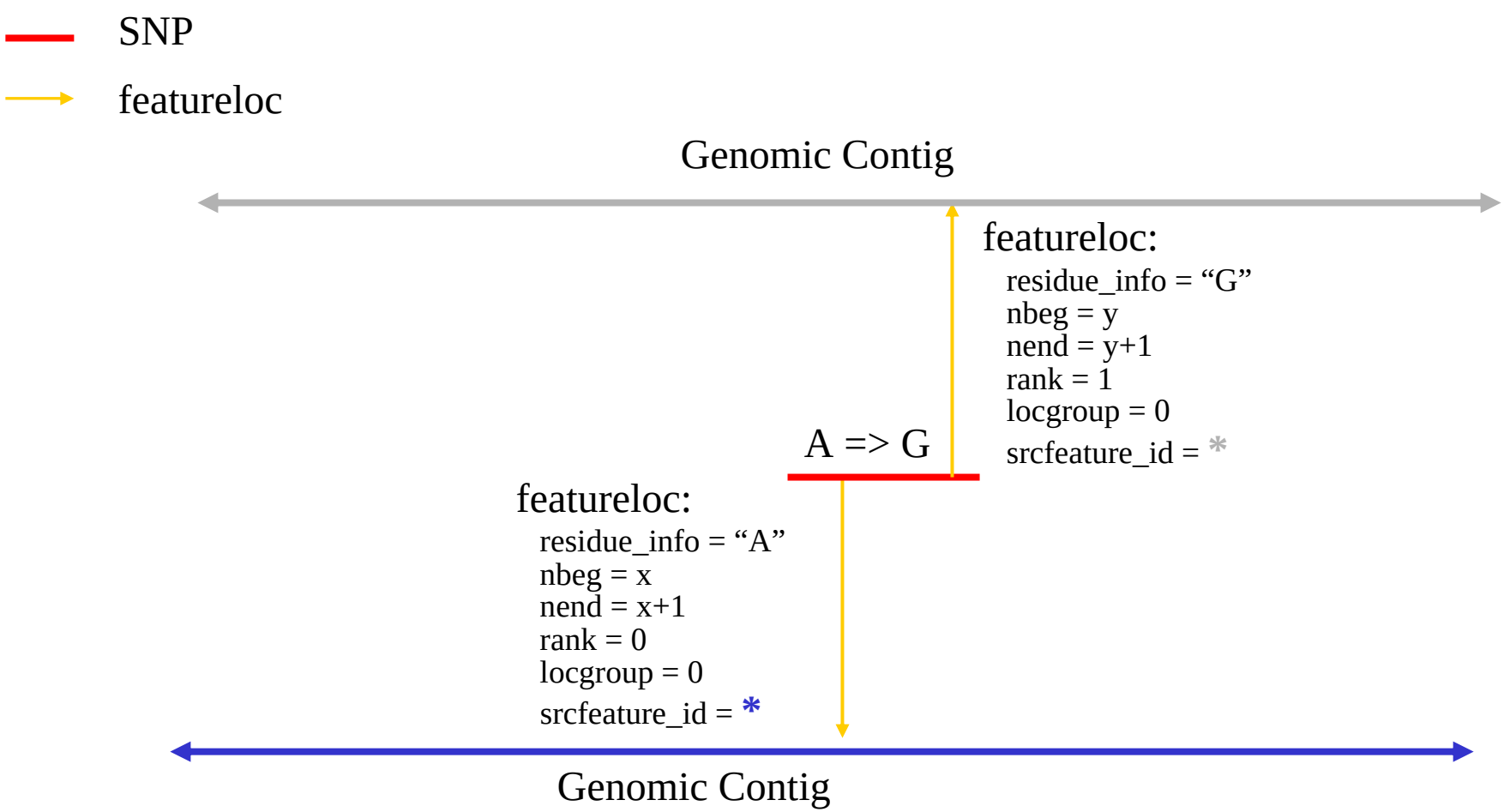
Bases | Arrows | Coverage



Capturing sequence variation data in Pathema

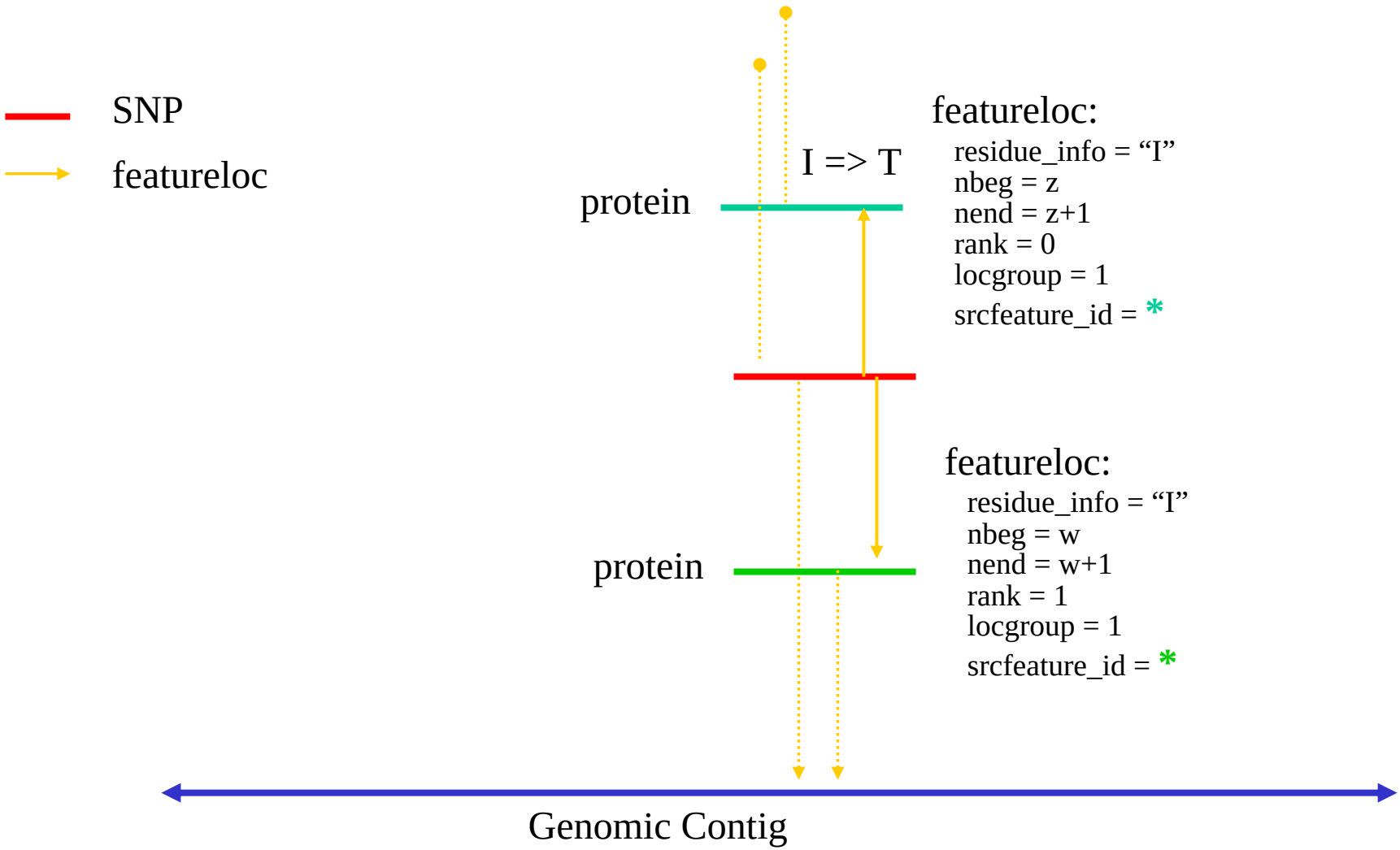
Sequence variation in Chado:

Single nucleotide polymorphisms



Sequence variation in Chado:

SNPs within protein coding regions



Prototype SNP report for *M. tuberculosis*

SNP report for reference sequence gmt_3810_assembly - Mozilla

File Edit View Go Bookmarks Tools Window Help

SNP report for re

Coord Loc Gene IDs, SNP location, amino acid code

SNP report for r

Please note: if "(-)" appears in the query position column, then the query base reported has been reverse complemented.

Ref asmbl	Ref pos	Ref type	Ref coding info	Query asmbl	Query pos	Query type	Query coding info	S/N
gmt_3810_assembly	1849 C	intergenic		bmt_689_assembly	11067 A	intergenic		
gmt_3810_assembly	11370 T	intergenic		bmt_689_assembly	20580 C	intergenic		
gmt_3810_assembly	11820 G	intergenic		bmt_689_assembly	21038 G	intergenic		
gmt_3810_assembly	14861 G	coding	ORF04265 MT0015 :: G (2)	bmt_689_assembly	24079 T	coding	ORF00026 :: V (2)	N
gmt_3810_assembly	15087 C	coding	ORF04267 MT0016 :: T (3)	bmt_689_assembly	24305 G	coding	ORF00028 :: T (3)	S
gmt_3810_assembly	15117 C	coding	ORF04267 MT0016 :: I (3)	bmt_689_assembly	24335 G	coding	ORF00028 :: M (3)	N
gmt_3810_assembly	16119 C	coding	ORF04269 MT0017 :: R (2)	bmt_689_assembly	25337 A	coding	ORF00027 :: L (2)	N
gmt_3810_assembly	17654 T	coding	ORF04270 MT0018 :: Q (2)	bmt_690_assembly	305 G	coding	ORF00030 :: P (2)	N
gmt_3810_assembly	17657 T	coding	ORF04270 MT0018 :: Q (2)	bmt_690_assembly	308 C	coding	ORF00030 :: R (2)	N
gmt_3810_assembly	25592 G	intergenic		bmt_690_assembly	8243 C	intergenic		
gmt_3810_assembly	27559 T	intergenic		bmt_690_assembly	10210 C	coding	ORF00038 :: L (1)	
gmt_3810_assembly	28260 C	coding	ORF04280 MT0026 :: G (3)	bmt_690_assembly	10911 T	coding	ORF00038 :: G (3)	S
gmt_3810_assembly	30500 T	coding	ORF04285 MT0029 :: G (3)	bmt_690_assembly	13151 C	coding	ORF00041 :: G (3)	S
gmt_3810_assembly	30669 G	coding	ORF04285 MT0029 :: A (1)	bmt_690_assembly	13320 T	coding	ORF00041 :: S (1)	N
gmt_3810_assembly	30924 T	coding	ORF04285 MT0029 :: S (1)	bmt_690_assembly	13575 C	coding	ORF00041 :: P (1)	N
gmt_3810_assembly	31058 T	intergenic		bmt_690_assembly	13709 C	intergenic		
gmt_3810_assembly	35552 C	coding	ORF04297 MT0037 :: A (3)	bmt_507_assembly	19286 (-)T	intergenic		
gmt_3810_assembly	35952 G	coding	ORF04297 MT0037 :: D (1)	bmt_507_assembly	18886 (-)C	intergenic		
gmt_3810_assembly	37249 C	coding	ORF04301 MT0040 :: S (2)	bmt_507_assembly	17589 (-)G	coding	ORF00052 :: W (2)	N
gmt_3810_assembly	37278 A	coding	ORF04301 MT0040 :: T (1)	bmt_507_assembly	17560 (-)T	coding	ORF00052 :: S (1)	N
gmt_3810_assembly	38974 C	intergenic		bmt_507_assembly	15864 (-)T	intergenic		
gmt_3810_assembly	39102 C	coding	ORF04302 MT0041 :: R (3)	bmt_507_assembly	15568 (-)G	coding	ORF00053 :: R (3)	S
gmt_3810_assembly	39197 C	coding	ORF04302 MT0041 :: G (1)	bmt_507_assembly	15473 (-)G	coding	ORF00053 :: R (1)	N
gmt_3810_assembly	40422 A	coding	ORF04303 MT0042 :: L (2)	bmt_507_assembly	14248 (-)C	coding	ORF00054 :: W (2)	N
gmt_3810_assembly	42225 C	coding	ORF04306 MT0044 :: C (2)	bmt_507_assembly	12445 (-)A	multiple		
gmt_3810_assembly	44712 A	coding	ORF04309 MT0047 :: R (1)	bmt_507_assembly	9958 (-)G	coding	ORF00059 :: G (1)	N
gmt_3810_assembly	48447 T	coding	ORF04313 MT0050 :: R (3)	bmt_507_assembly	6223 (-)C	coding	ORF00062 :: R (3)	S
gmt_3810_assembly	49304 C	coding	ORF04314 MT0051 :: V (1)	bmt_507_assembly	5366 (-)T	coding	ORF00063 :: I (1)	N
gmt_3810_assembly	51593 C	coding	ORF04318 MT0053 :: V (1)	bmt_507_assembly	3079 (-)G	coding	ORF00065 :: L (1)	N
gmt_3810_assembly	51898 A	coding	ORF04320 MT0054 :: D (3)	bmt_507_assembly	2774 (-)T	multiple		
gmt_3810_assembly	53729 G	coding	ORF04322 MT0056 :: M (3)	bmt_507_assembly	943 (-)C	intergenic		
gmt_3810_assembly	57337 T	coding	ORF04325 MT0058 :: F (2)	bmt_506_assembly	2579 A	coding	ORF00072 :: Y (2)	N
gmt_3810_assembly	62601 A	coding	ORF04335 MT0064 :: P (3)	bmt_486_assembly	284 G	intergenic		
gmt_3810_assembly	63090 G	intergenic		bmt_486_assembly	773 T	intergenic		
gmt_3810_assembly	65094 C	coding	ORFD0005 MT0066.1 :: W (3)	bmt_486_assembly	2777 T	intergenic		

http://manatee.tigr.org/tigr-scripts/sybil/showsnp.cgi?db=mycotb...f_seq_id=gmt_3810_assembly&ref_posn=30669&left_bp=24&right_bp=24

SNP report (preliminary design)

SNP: reference=gmt_3810_assembly [30669] - Mozilla

File Edit View Go Bookmarks Tools Window Help

SNP report for reference sequence g... | SNP: reference=gmt_3810_assembly ...

```
                                  S
                                  N
                                  P
                                  |
                                  |
   +1 R  D  P  A  D  Y  E  H  A  N  E  P  D
   +1 R  D  P  A  D  Y  E  H  A  N  E  P  D      gmt.ORF04285_cds;MT0029/+ strand
   +2  G  I  R  P  I  M  S  M  P  T  N  R
   +3   G  S  G  R  L  *  A  C  U  R  T  G
30644 AGGGATCCGGCCGATTATGAGCATGCCAACGAACCGGATGAGCGTCGCG 30693      gmt_3810_assembly/+ strand
   -1    P  D  P  R  N  H  A  H  W  R  V  P  H  A  D  R
   -2   P  I  R  G  I  I  L  M  G  V  F  R  I  L  T  A
   -3     S  G  A  S  *  S  C  A  L  S  G

   +1 R  D  P  A  D  Y  E  H  S  N  E  P  D
   +1 R  D  P  A  D  Y  E  H  S  N  E  P  D      ntmb01.NTORF0027_cds;NTL01MB0027/+ strand
   +2  G  I  R  P  I  M  S  I  P  T  N  R
   +3   G  S  G  R  L  *  A  F  U  R  T  G
30643 AGGGATCCGGCCGATTATGAGCATTCCAACGAACCGGATGAGCGTCGCG 30692      ntmb01_2_assembly/+ strand
   -1    P  D  P  R  N  H  A  N  W  R  V  P  H  A  D  R
   -2   P  I  R  G  I  I  L  M  G  V  F  R  I  L  T  A
   -3     S  G  A  S  *  S  C  E  L  S  G

   +1 R  D  P  A  D  Y  E  H  S  N  E  P  D
   +1 R  D  P  A  D  Y  E  H  S  N  E  P  D      bmt.ORF00041_cds;/+ strand
   +2  G  I  R  P  I  M  S  I  P  T  N  R
   +3   G  S  G  R  L  *  A  F  U  R  T  G
13295 AGGGATCCGGCCGATTATGAGCATTCCAACGAACCGGATGAGCGTCGCG 13344      bmt_690_assembly/+ strand
   -1    P  D  P  R  N  H  A  N  W  R  V  P  H  A  D  R
   -2   P  I  R  G  I  I  L  M  G  V  F  R  I  L  T  A
   -3     S  G  A  S  *  S  C  E  L  S  G  S  S  R  R

      AGGGATCCGGCCGATTATGAGCATTCCAACGAACCGGATGAGCGTCGCG

      0000000000000000000000000000000000000000000000000
      5545555555555554444444444444444444444444445555

      2222233233332222322333333333333333333333333333223
      8897900912118769097111111201125344323233643321990

   +1 R  D  P  A  D  Y  E  H  S  N  E  P  D  E  R  R
   +1 R  D  P  A  D  Y  E  H  S  N  E  P  D  E  R  R
   +2  G  I  R  P  I  M  S  I  P  T  N  R  M  S  V  A
   +3   G  S  G  R  L  *  A  F  Q  R  T  G  *  A  S
30663 AGGGATCCGGCCGATTATGAGCATTCCAACGAACCGGATGAGCGTCGCG 30712      ntmt02_1_assembly/+ strand
   -1    P  D  P  R  N  H  A  N  W  R  V  P  H  A  D  R
   -2   P  I  R  G  I  I  L  M  G  V  F  R  I  L  T  A
   -3     S  G  A  S  *  S  C  E  L  S  G  S  S  R  R

                                  |
                                  |
                                  S
                                  N
                                  P
```

ATGCC

ATTCC

ATTCC

Depth of coverage

Consensus quality

Click here to show the overlapping exons/ORFs in their entirety

Indel report (preliminary design)

Indel: indel_id=91824_insertion - Mozilla

File Edit View Go Bookmarks Tools Window Help

Indel: indel_id=91824_insertion

Logged into [mycotb] as crabtree
Home

```
1034199 TGGTGGAGCTGATGTTGATCACGCTGCCGCCGCCGGAGTGTTCCAGCATCAA 1034251      gmt_3810_assembly/+ strand
        |||||||||||||||||||||||   |||||||||||||||||||||||||
  29863 TGGTGGAGCTGATGTTGATCACGC---CGCCGCCGGAGTGTTCCAGCATCAA 29814        bmt_685_assembly/- strand

        TGGTGGAGCTGATGTTGATCACGC   CGCCGCCGGAGTGTTCCAGCATCAA              al

        000000000000000000000000   000000000000000000000000
        555555555544444444444444   444444444444444444444444

        223332222333333333322323   332322322333333333333333
        951108889022124501078184   418188186102135445422315 3
```

read_depth [bmt_685_assembly min=1 max=15 avg=6]

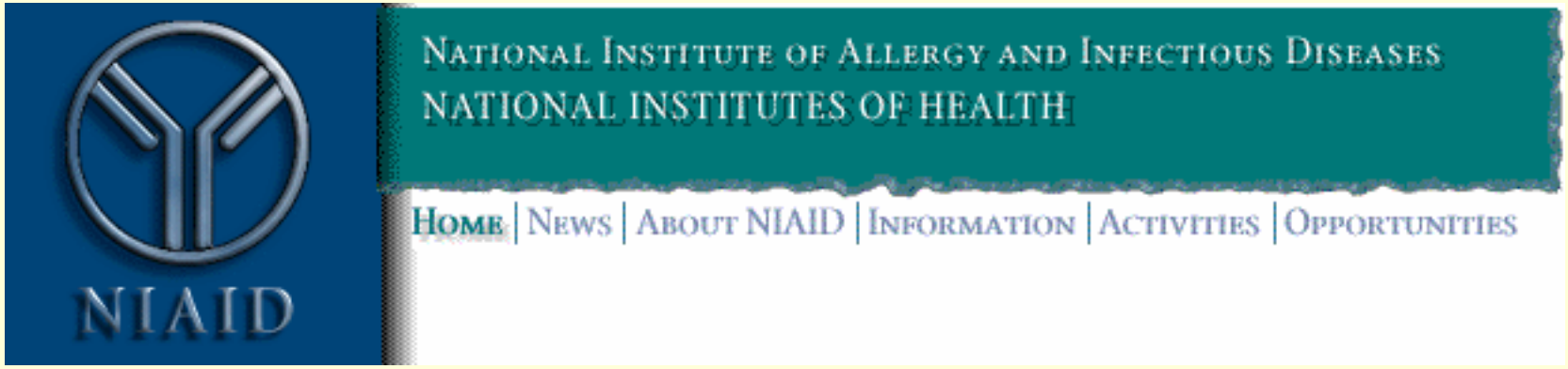
co 4 max=39 avg=31]

Depth of coverage

Consensus quality

Feedback

Done

The Assembly and Polymorphism Team

- Art Delcher
- Michael Schatz
- Adam Phillippy
- Mihai Pop
- Sam Angiuoli
- Jonathan Crabtree (web)
- Jay Sundaram (Chado)
- Dave Emmert at Flybase (Chado database)
- Jeremy Peterson
- Owen White