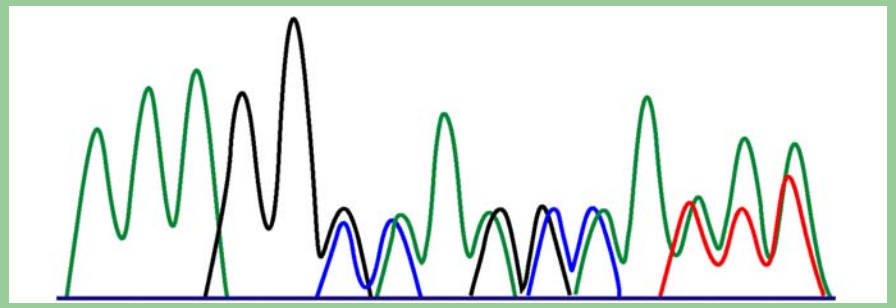
Decoding of Superimposed Traces Produced by Direct Sequencing of Heterozygous Indels



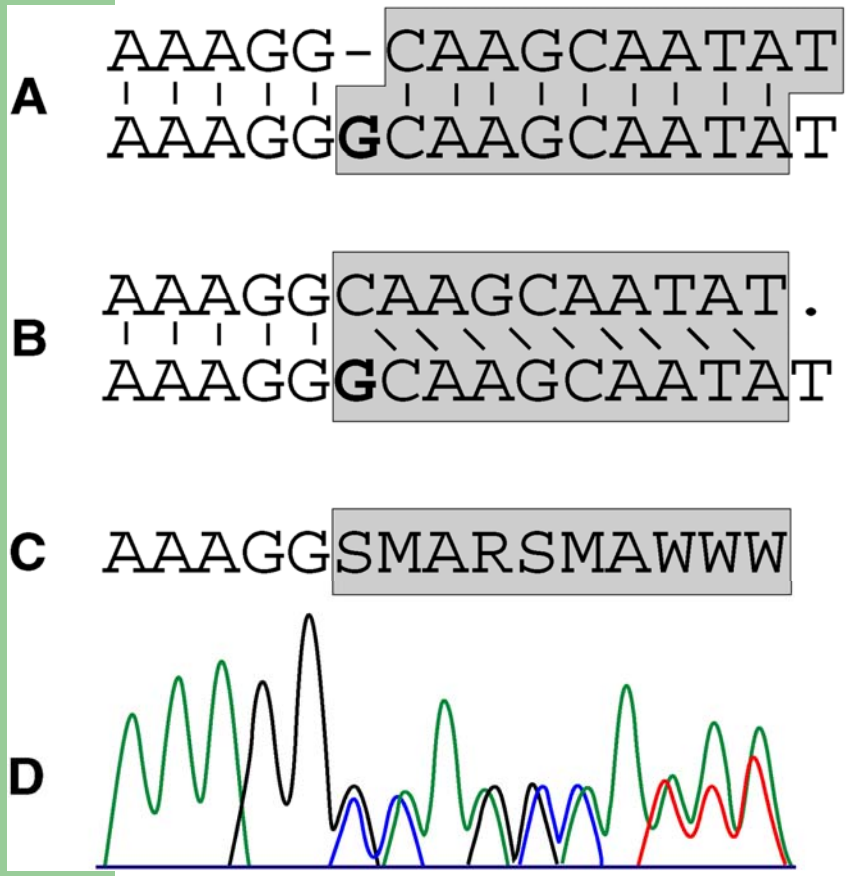
Dmitriev, D.A. & Rakitov, R.A.

Illinois Natural History Survey,
Institute of Natural Resource Sustainability,
University of Illinois at Urbana-Champaign,
1816 S. Oak st.,
Champaign IL, 61820
http://ctap.inhs.uiuc.edu/dmitriev/indel.asp

ILLINOIS NATURAL HISTORY SURVEY

1867

Problem of double peaks on chromatogram



IUPAC symbols:
R = A or G
Y = C or T
K = G or T
M = A or C
S = C or G
W = A or T
B = not A
D = not C
H = not G
V = not T
N = any

A pair of allelic sequences properly aligned (**A**), unaligned (**B**), translated into a consensus (**C**), and resulted chromatogram (**D**).

The following applications could be used to call double paeks PHRED, KB Basecaller, Sequencher

Reasons for double peaks

- Direct sequencing of diploid alleles containing heterozygous insertions/deletions. (Mixed trace downstream of the indel is formed by two allelic traces superimposed onto each other with a phase shift).
- Sequencing of unrelated templates and alternative splicing.
- Single nucleotide polymorphisms (SNPs) or base calling errors due to low quality of chromatogram (individual double picks)

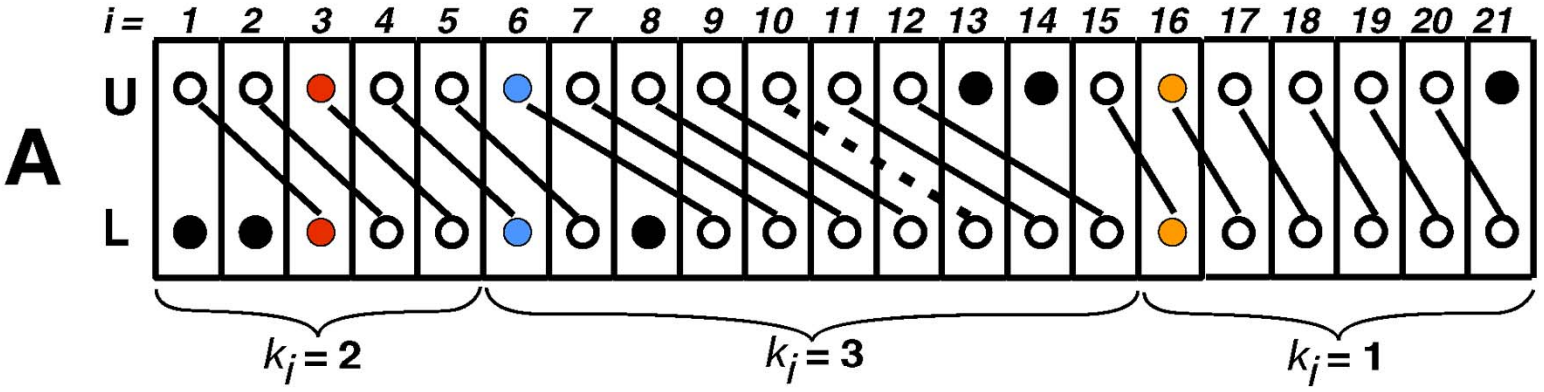
Solutions

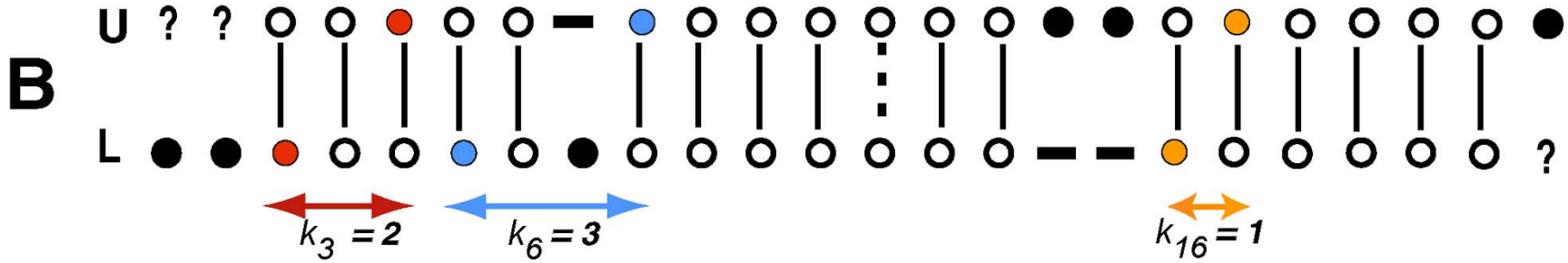
- Discard as uninterpretable.
- Use new sequencing technologies, such as pyrosequencing, which works with single DNA molecules.
- Separating the templates prior to sequencing via cloning into a vector or selectively amplifying one allele using allele-specific primers.
- Computational methods to extract information from mixed traces.

Computational methods for decoding of mixed traces

- **Subtracting a reference sequence**: PolyPhred, STADEN, CodonCode Aligner, Mutation Surveyor, InSNP, PolyScan, AutoCSA, and Tenney, A.E at al. (2007) application to automatically resolving double traces aligning the mixed sequence to genomic database
- **Using reverse sequence as a reference**: SeqScape, Varian Reporter, Champuru.
- **Extracting information from an individual mixed trace (without reference)**: Shift Detector, CodonCode Aligner Ver. 2, In*d*elligent.

Phase shift and two renderings of the same alignment. Optimality criterion.

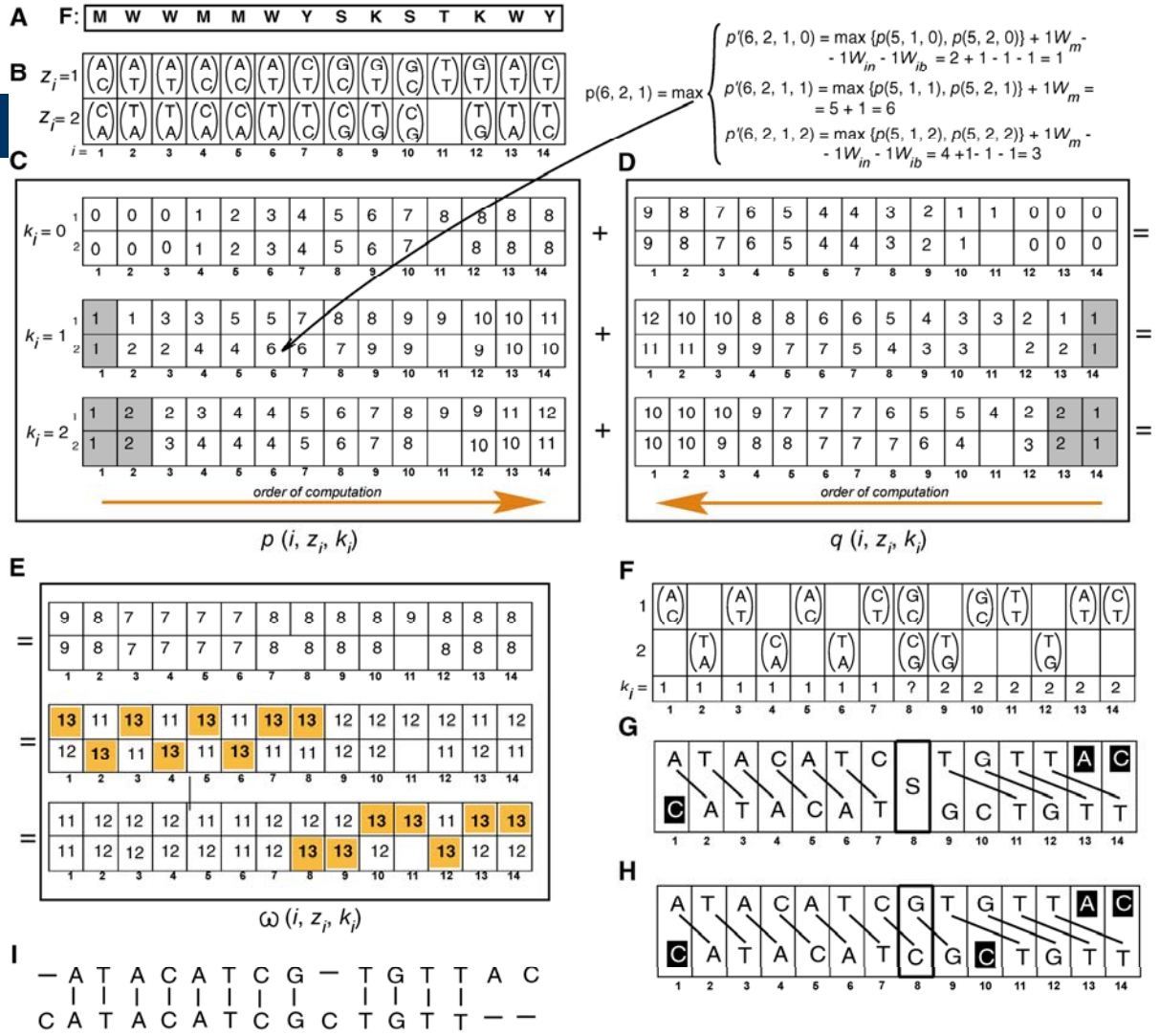




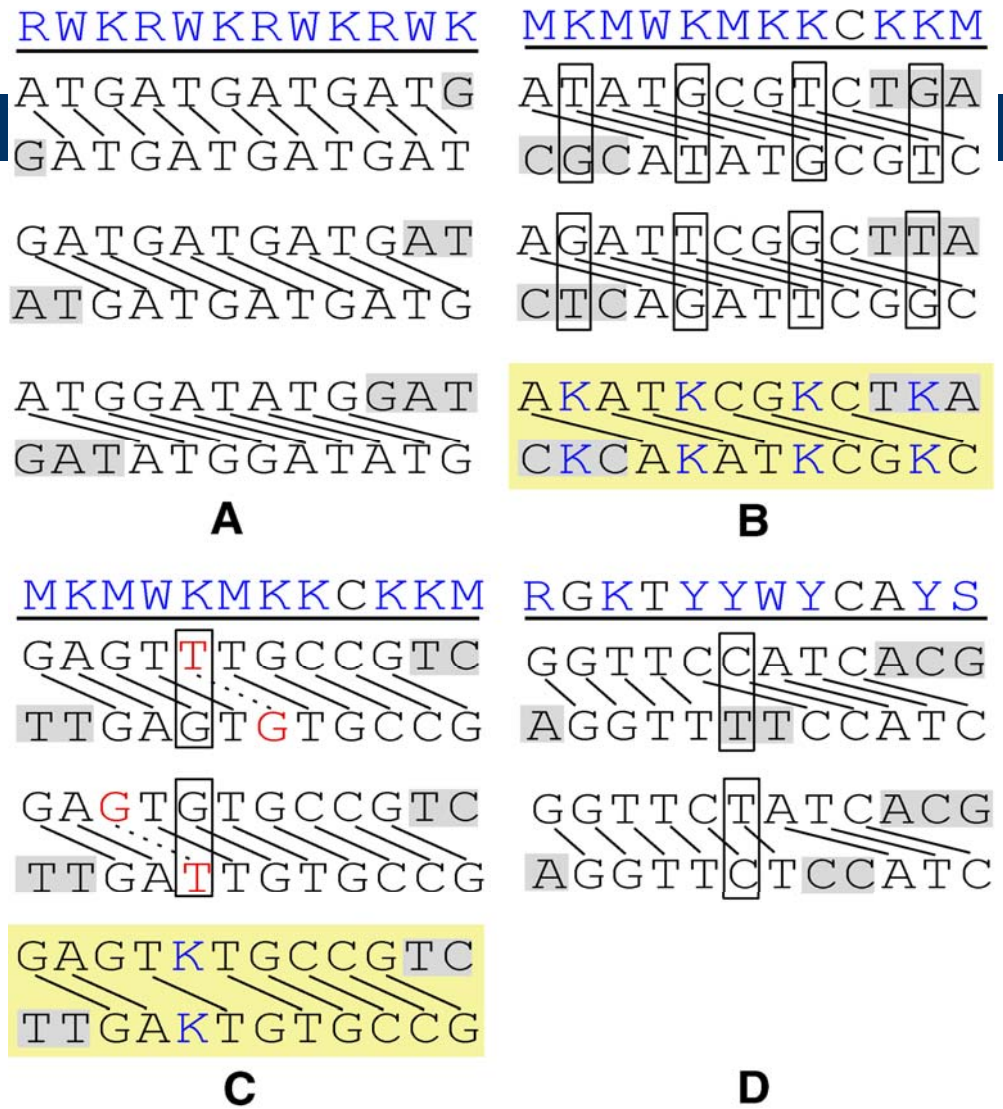
V=n - (#mismatches) - (#insertions) - (#inserted bases)

Number of solutions = $2^{n-1} \times k_{max}$

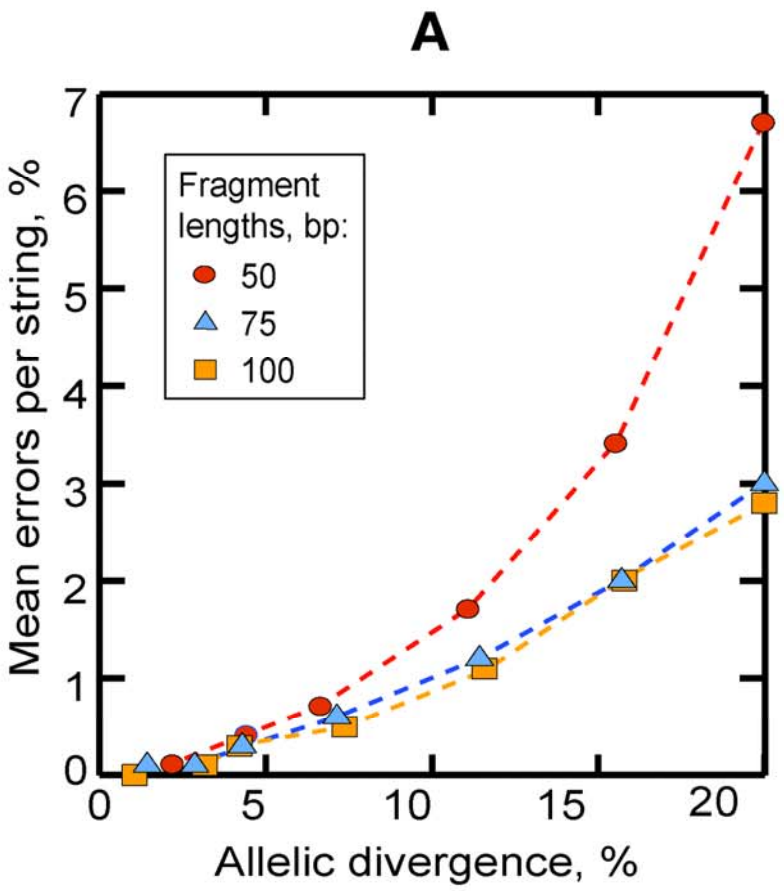
Dynamic optimization algorithm

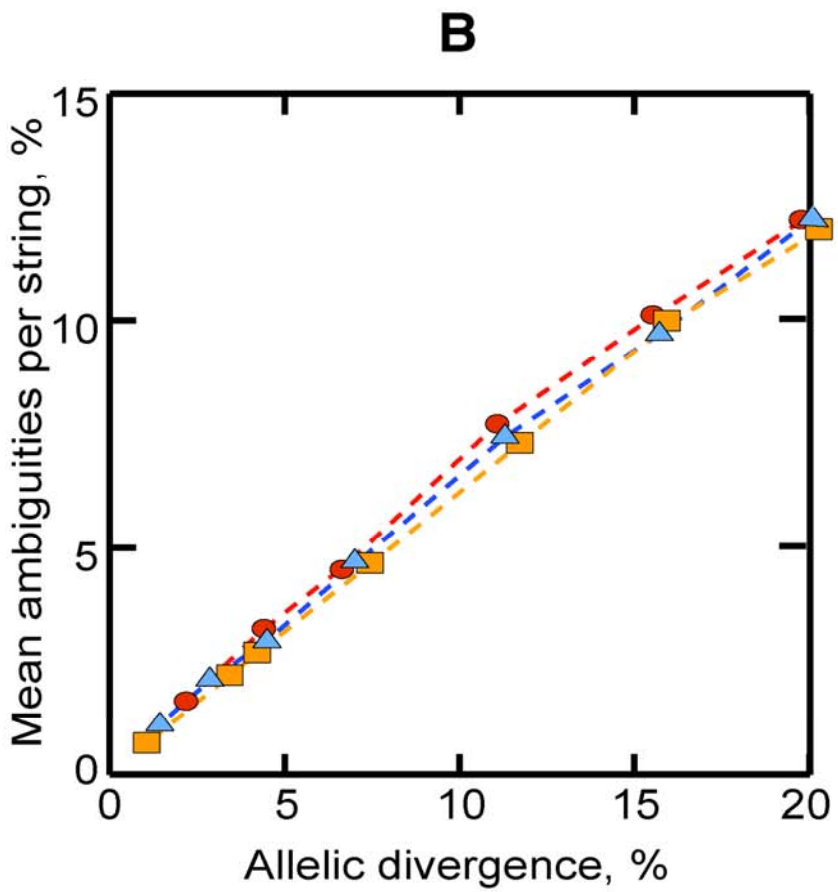


Multiple cooptimal solutions

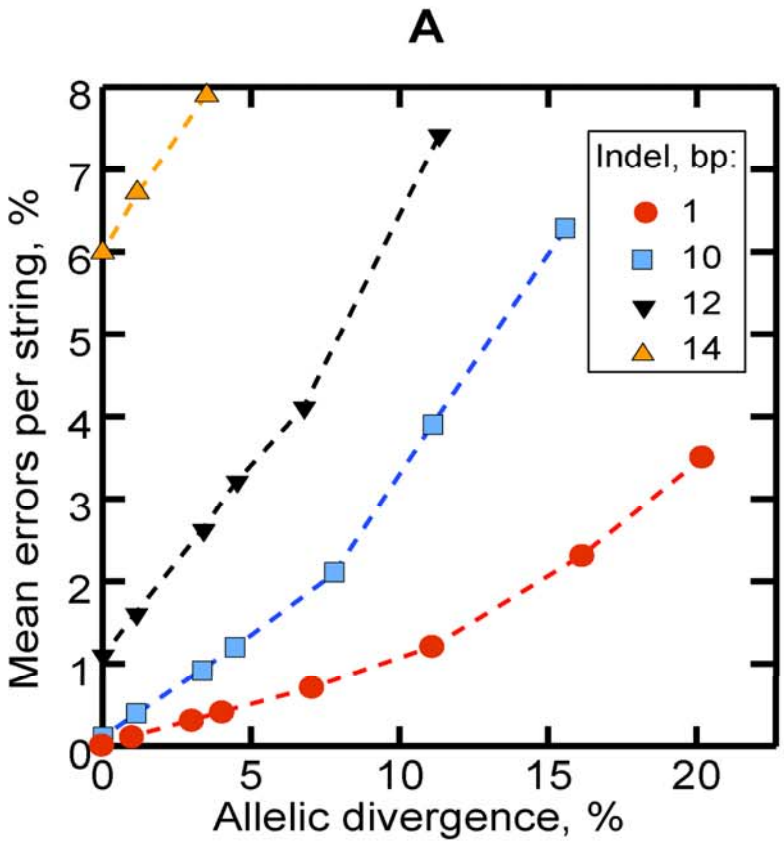


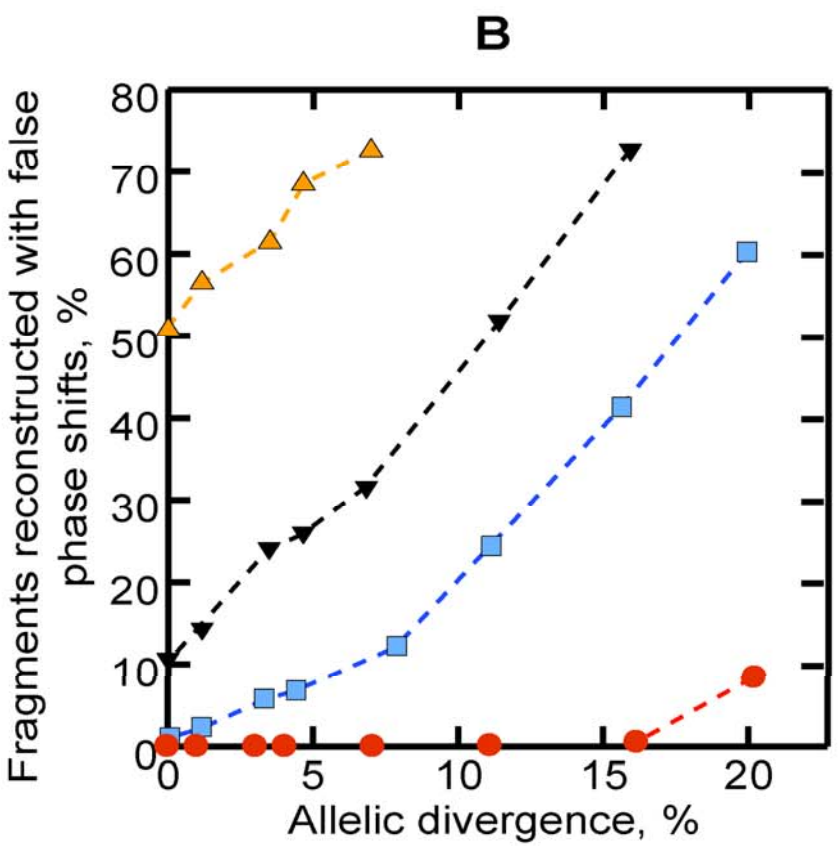
Accuracy of decoding of simulated mixed fragments formed by a single 5bp phase shift (1000 runs for each point)



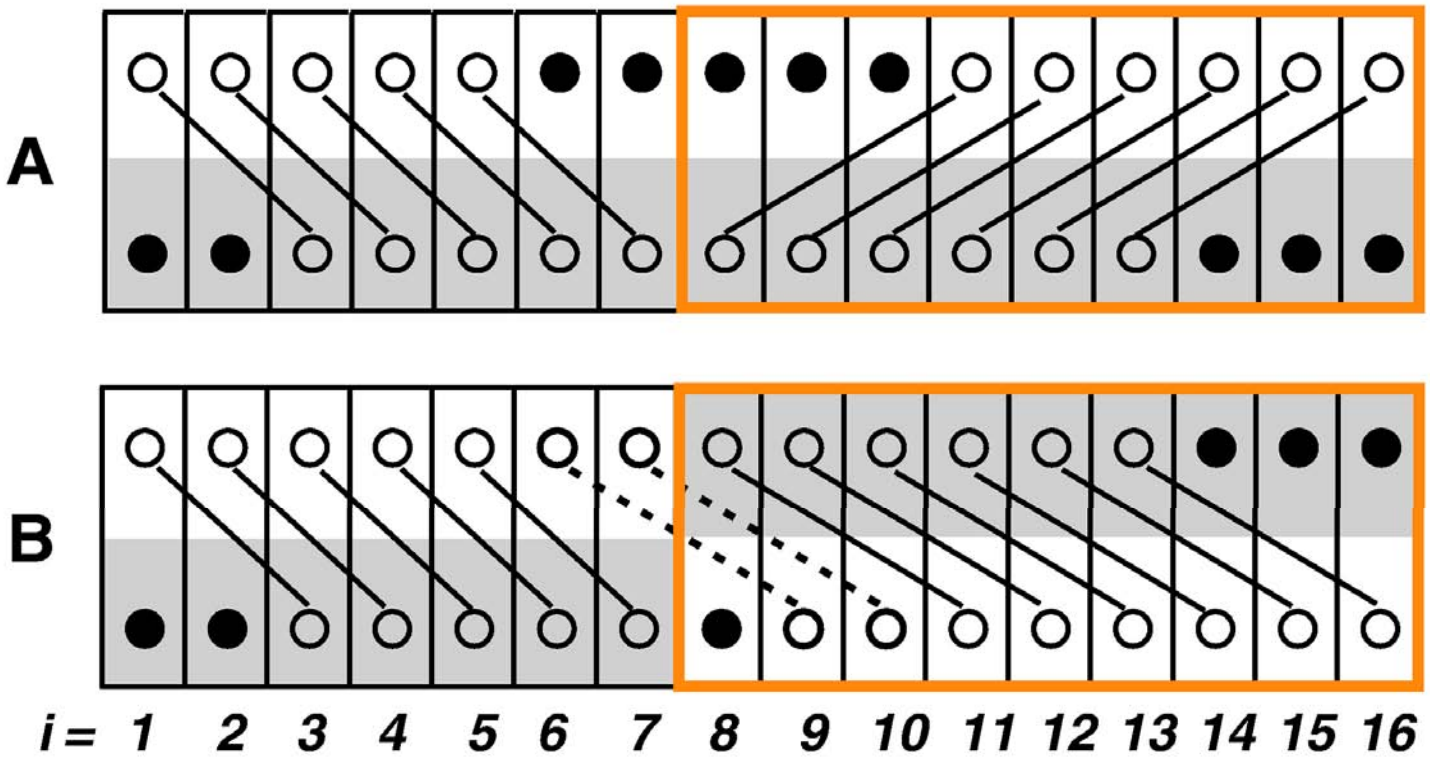


Accuracy of decoding of simulated mixed 100 bp fragments formed with an insertion of variable size in the middle (1000 runs for each point)





Two aligned solutions of the same mixed fragment representing the transition between two phase sifts (“Long” vs. “Short” indel)



Validation with human traces

- We used 104 (103-677bp, with indels 5-30 bp) traces of 198 recorded by Bhangale et al. (2005) from NCBI Trace Archive as having heterozygous indel.
- Sequencer was used to call second paeks.
- After reconstruction, traces were aligned with best matching human sequnces in NCBI Trace Archive.
- 102 traces reconstructed with a single indel, two with two indels.
- 67 traces without errors, 31 – with 1-2 errors, 6 – with 3-7 errors. Half of the fragments reconstructed without ambiguities.
- Mean of 99.1±1.25% of bases per fragment decoded correctly and unambiguously (in the same conditions ShiftDetector decoded only 72.5±6.47% of bases).
- About 60% of reconstructed mean of 0.66 errors per fragment were due base calling errors, mostly in low-quality trace regions.

Conditions for reconstructions with 99.5-100% accuracy

- Homologous fragments resulted from indel mutation.
- Analyzed fragment is significantly larger than the indel (at least 10 times; in human 92.3% of indels are 1-10bp).
- Low divergence between mixed traces (<5%; for human noncoding DNA, the average divergence is <0.1%, fruit fly 1-2%, sea squirts 4.5%).
- Multiple indels, if present, are well spaced.
- Methods relies on base calling software.

Indelligent interface

Reconstruction results

IUPAC sequence

GTTCCGGATTACCTTTTTAGTGGCTGGTGACGACAACATCGGCAGTGGGGRSRYGMYRYKYKRSYYGWWCMRYTMRRSKMGGGKSKKRWGKWRRRGRRRM

Reconstructed sequences

GTTCCGGATTACCTTTTTAGTGGCTGGTGACGACAACATCGGCAGTGGGG.....ACATGCTGCTTGAGCCGTTCAACTAGGGGCGGGTGTTAAGGAGGAGGAAC

GTTCCGGATTACCTTTTTAGTGGCTGGTGACGACAACATCGGCAGTGGGGGGGCGACATGCTGCTTGAACCGTTCAACTAGGGGCGGGTGTTAAGGAGGA.....

Combined sequence

GTTCCGGATTACCTTTTTAGTGGCTGGTGACGACAACATCGGCAGTGGGGGGGCGACATGCTGCTTGARCCGTTCAACTAGGGGCGGGTGTTAAGGAGGAGGAAC

Statistics

Length of original sequence: **100**.

Number of resolved positions: **94** (94 %).

Number of mismatches: **1** (1 %).

Number of ambiguities: **1** (1 %).

Percent of resolved ambiguities: **97.56** %. (Maximum expected by chance: 80.07 %).

A floating indel has been reconstructed.

Detected phase shifts:

0 bp., 50 positions.

5 bp., 50 positions.

Script running time: 0.703125 sec.

Number of users: **205**.

Fragments analyzed: **7618** (**2916260** bp).

Acknowledgments

We are thankful to Saurabh Sinha for valuable discussion and suggestions

Tushar Bhangale for providing information which enable us to obtain human traces for testing.

Chris Dietrich for helpful comments and support.

The work was partially supported by NSF grants.