Sequence-specific error profile of Illumina sequencers

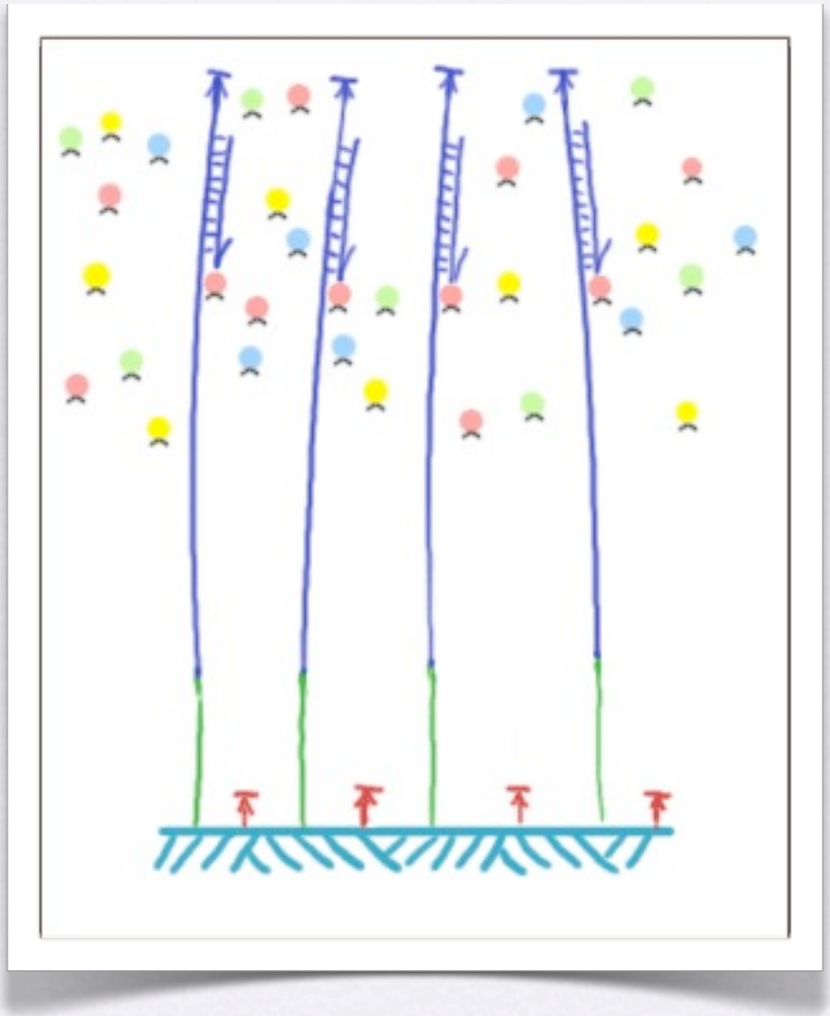
Kensuke Nakamura[1,*], Taku Oshima[2], Takuya Morimoto[2,3], Shun Ikeda[1], Hirofumi Yoshikawa[4,5], Yuh Shiwa[5], Shu Ishikawa[2], Margaret C. Linak[6], Aki Hirai[1], Hiroki Takahashi[1], Md. Altaf-Ul-Amin[1], Naotake Ogasawara[2] and Shigehiko Kanaya[1]

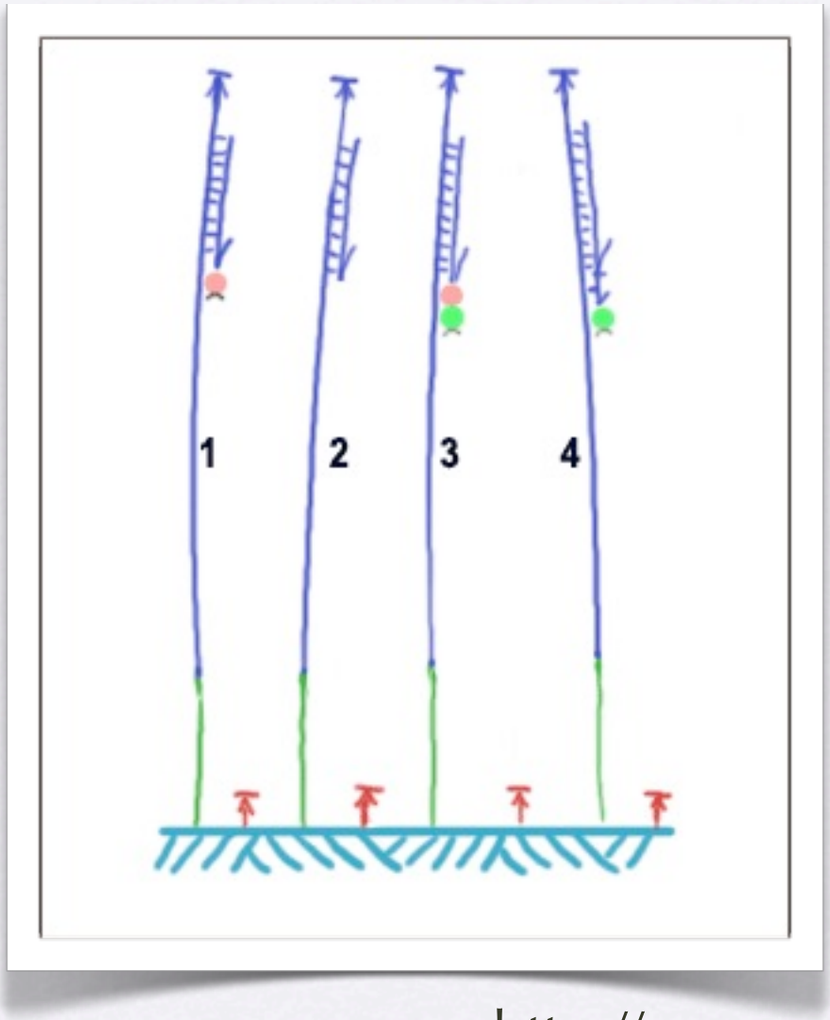
[1]Graduate School of Information Science, [2]Graduate School of Biological Sciences, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan, [3]Biological Science Laboratories, Kao Corporation, 2606 Akabane, Ichikai, Haga, Tochigi 321-3497, [4]Department of Bioscience, Tokyo University of Agriculture, [5]Genome Research Center, NODAI Research Institute, Tokyo University of Agriculture, 1-1-1 Sakuragaoka Setagaya-ku, Tokyo, 156-8502, Japan and [6]Department of Chemical Engineering and Material Science, University of Minnesota, 223 Amundson Hall, 421 Washington Avenue S.E., Minneapolis, MN 55455, USA

Received February 3, 2011; Revised April 25, 2011; Accepted April 26, 2011

Andres Veidenberg
JClub 13.06.2011

Illumina sequencing





http://seq.molbiol.ru

1: normal extension -> no dephasing
2: nucleotide not incorporated -> negative dephasing on next cycle
3: incorporated nucleotide has no terminator -> positive dephasing
4: incorporated nucleotide has no fluorophore and terminator -> positive dephasing

Illumina error profiles

- Coverage variation
  - bias of PCR - secondary structures in ssDNA
  - lower coverage in AT-repeats
- Miscalls
  - more substitution-type than indel-type miscalls
  - more frequent in first and last cycles
  - more frequent in GC-rich regions
  - A->C and C->G are observed more often than others
  - Read quality significantly lower in later cycles (lagging-strand dephasing)

Sequence specific errors

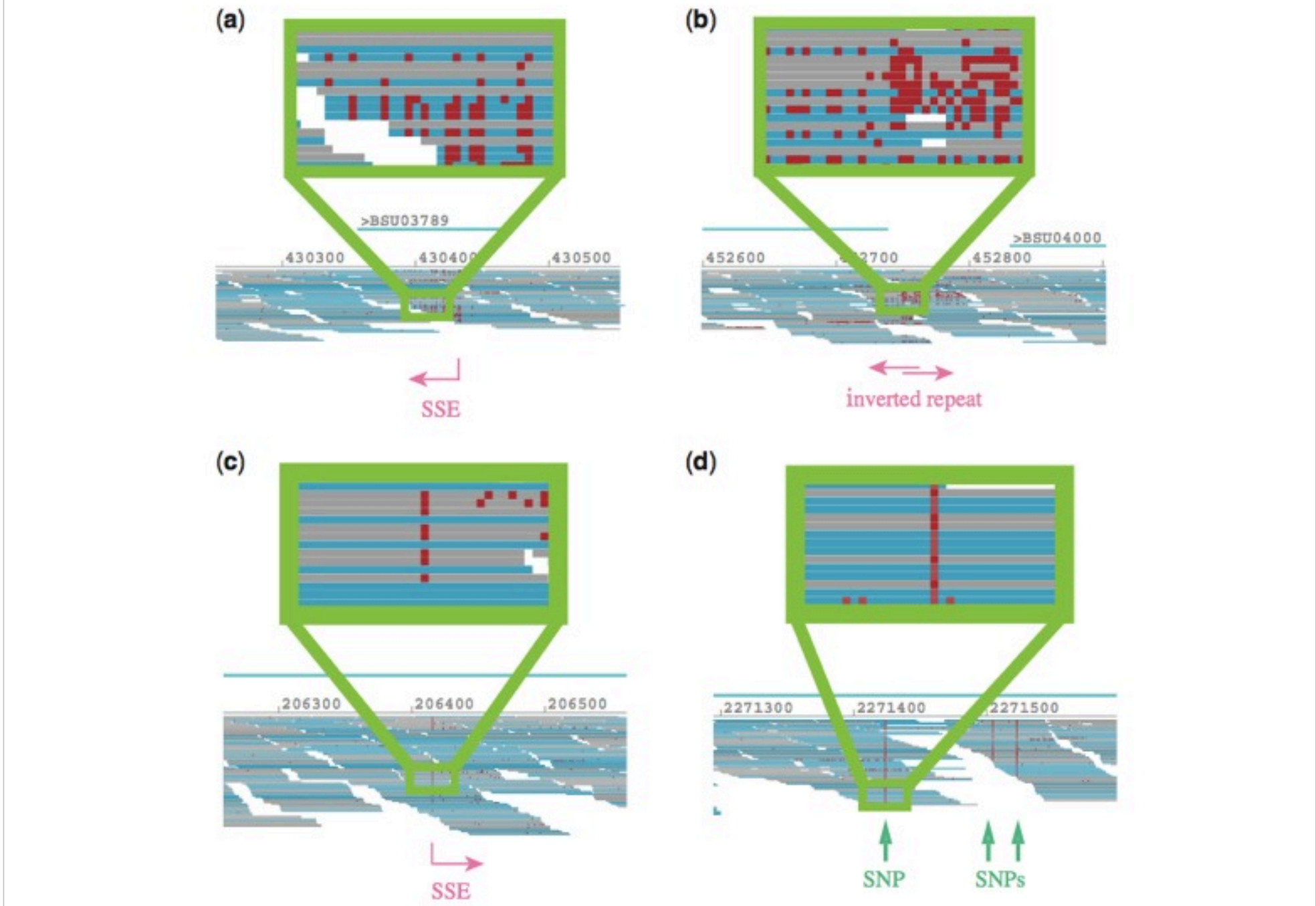


Figure 2. Examples of SSE and SNP positions in mapping of *B. subtilis*. Each drawing displays areas with (**a**) an SSE position, (**b**) two overlapping SSE positions with inverted repeat, (**c**) an SSE resembling an SNP and (**d**) true SNPs.

SSE positions

- Searching for sequence specific error positions
  - mismatches in >30% of reads in same direction
  - 4 other such mismatches in 40bp downstream
  - no mismatches in 40bp upstream

SSE positions

- 547 SSE positions identified in *B. subtilis* genome
  - (C/T)GGC(G/T) in most SSE positioins
- Some SSE positions close to inverted repeats

Table 2. Number of SSE positions detected automatically

Species	Forward	Backward	Total	Ref. length	SSE occurrence (one per bp)	GC contents (%)
Bacillus subtilis	287	287	574	215 606	7344	43.5
Mycobacterium bovis	4374	4273	8647	4 345 492	502	65.4
Staphylococcus aureus	353	329	682	2 903 081	4256	32.7
Bordetella pertussis	2747	2675	5422	4 086 189	754	67.7

SSE mismatch patterns

- A mismatched base in SSE region was often similar to a preceding reference base

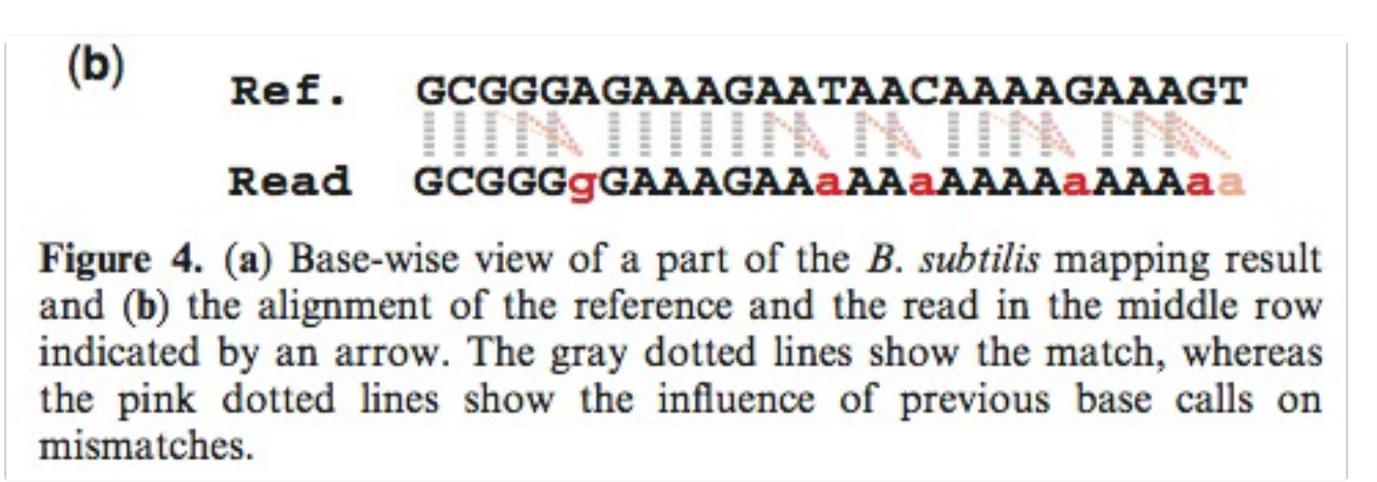


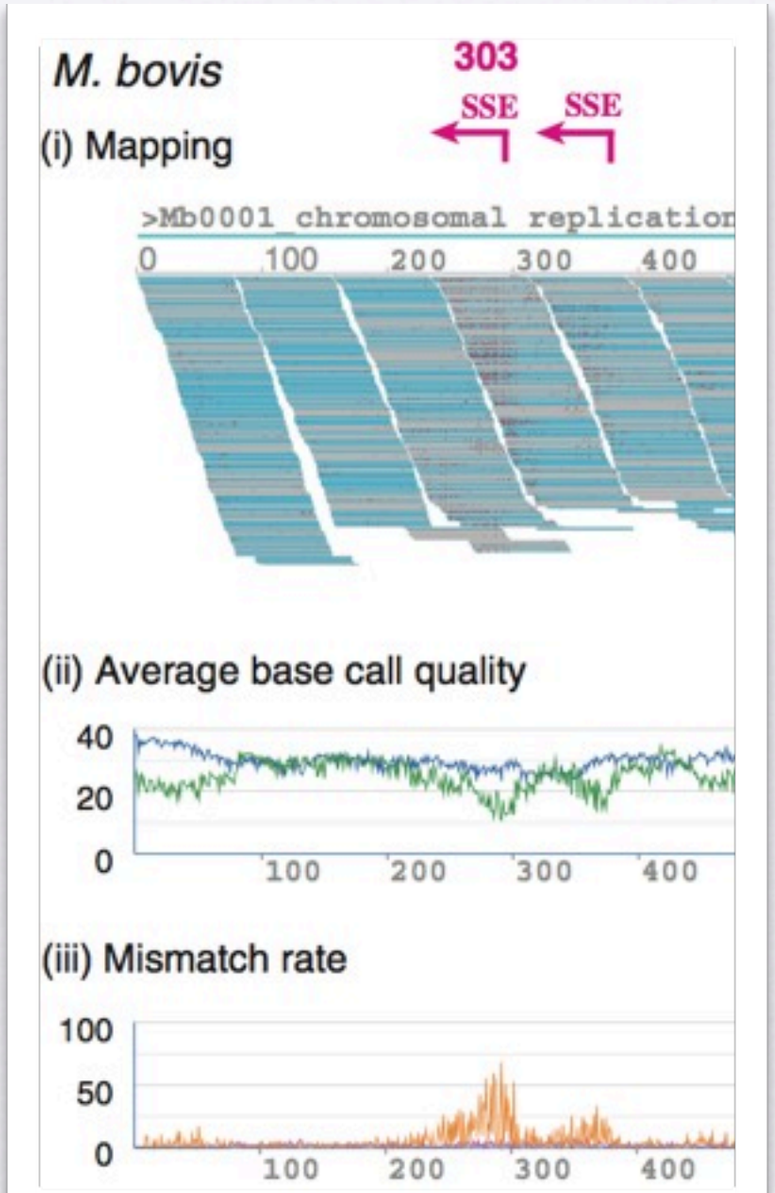
Figure 4. (**a**) Base-wise view of a part of the *B. subtilis* mapping result and (**b**) the alignment of the reference and the read in the middle row indicated by an arrow. The gray dotted lines show the match, whereas the pink dotted lines show the influence of previous base calls on mismatches.

- SSE mismatch conversion is influenced by GC content

SSE mismatches

- Particular sequence positions are associated with low base-call quality and high mismatch rate

- Miscalls in sequencing/base call quality ratio get worse in later cycles -> SSE comes more evident with increasing Illumina sequencers read length



SSE mechanisms

- Long inverted repeat enhances folding of ssDNA
- Similarity between GGC/inverted repeats mismatch patterns suggest the same mechanism
- Preference of DNA polymerase is most likely to be responsible

Mechanism: inverted repeats

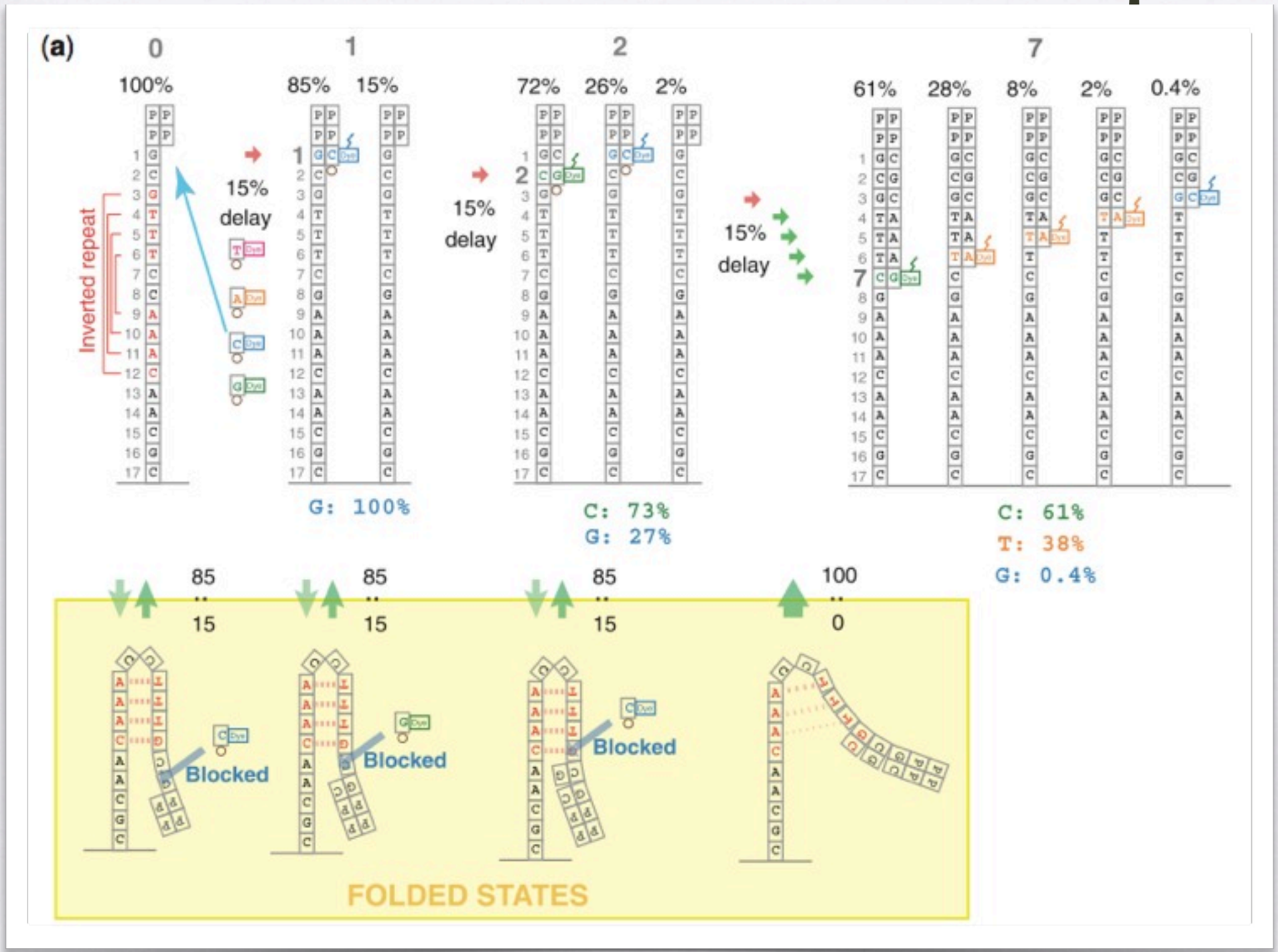


Figure 6. Schematic representation of the (**a**) inverted repeat and (**b**) enzyme preference for the SSE hypothetical mechanistic models. The gray numbers at the top indicate the cycle number and the numbers below indicate the relative population of each single-stranded DNA during the cycle. The colored bases and numbers below the drawings show the relative intensity of signals during that cycle. For instance, the second cycle of model (a) emits signals for C and G with an intensity of 73 and 27%, respectively.

Mechanism: GGC sequence

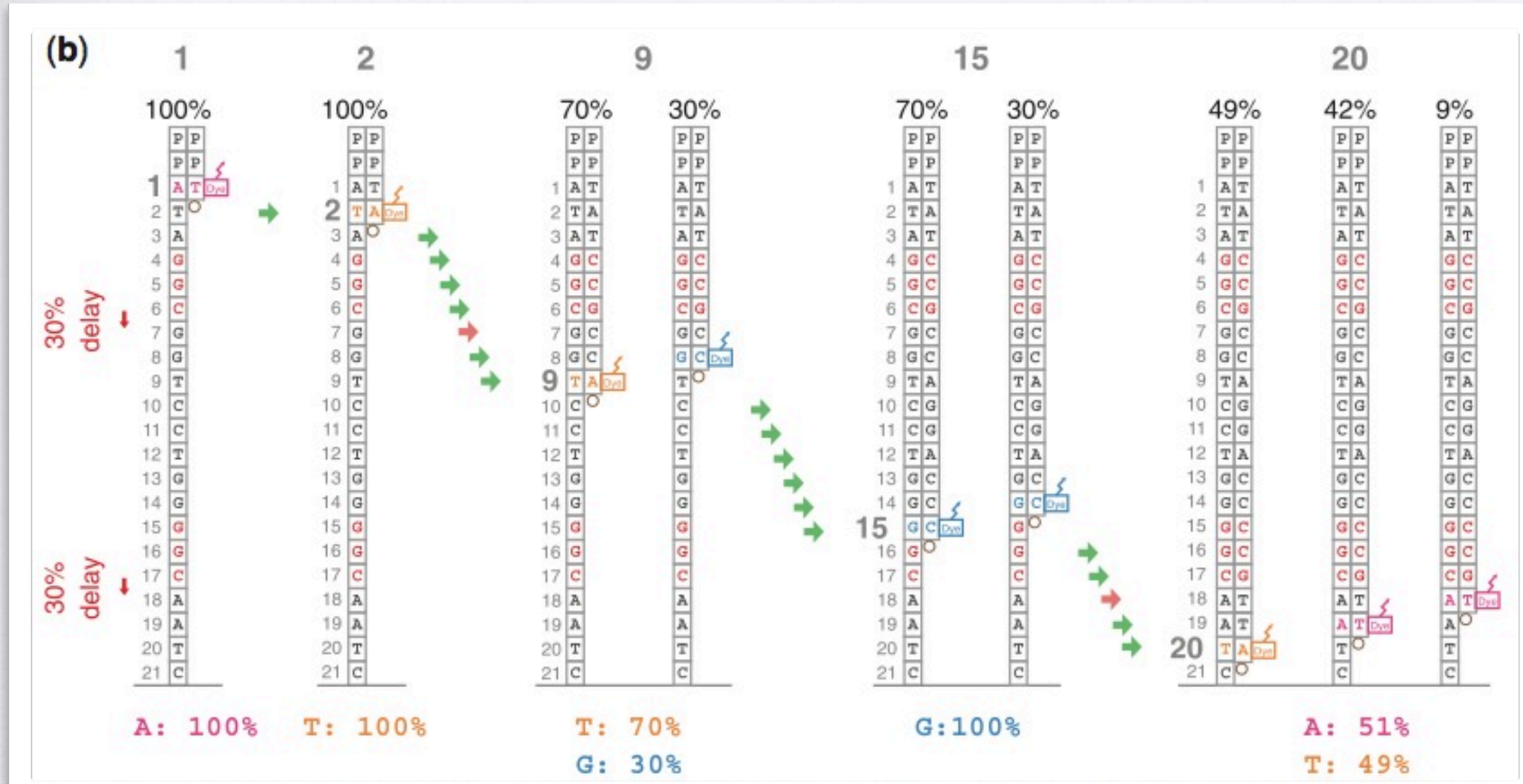
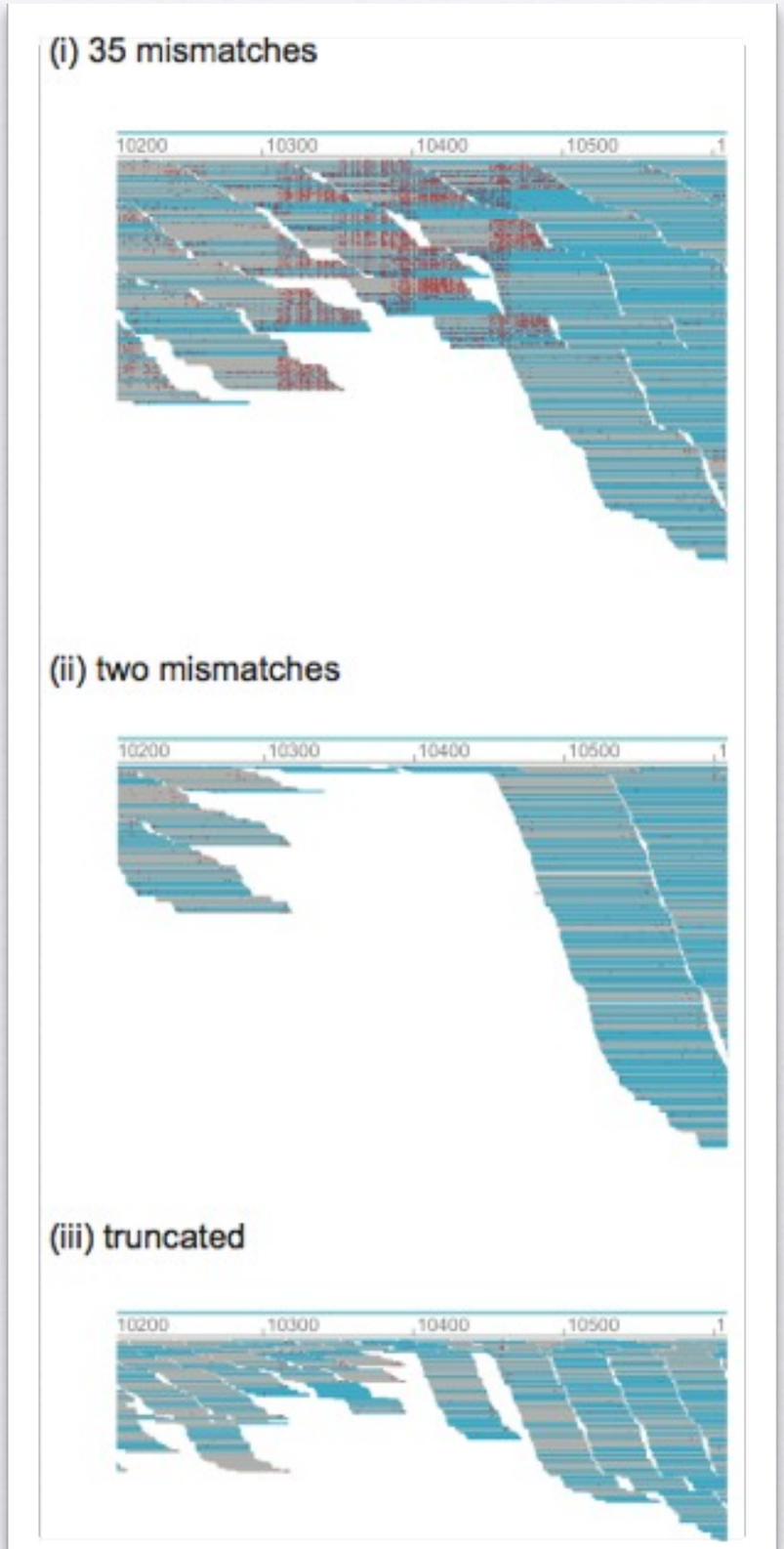
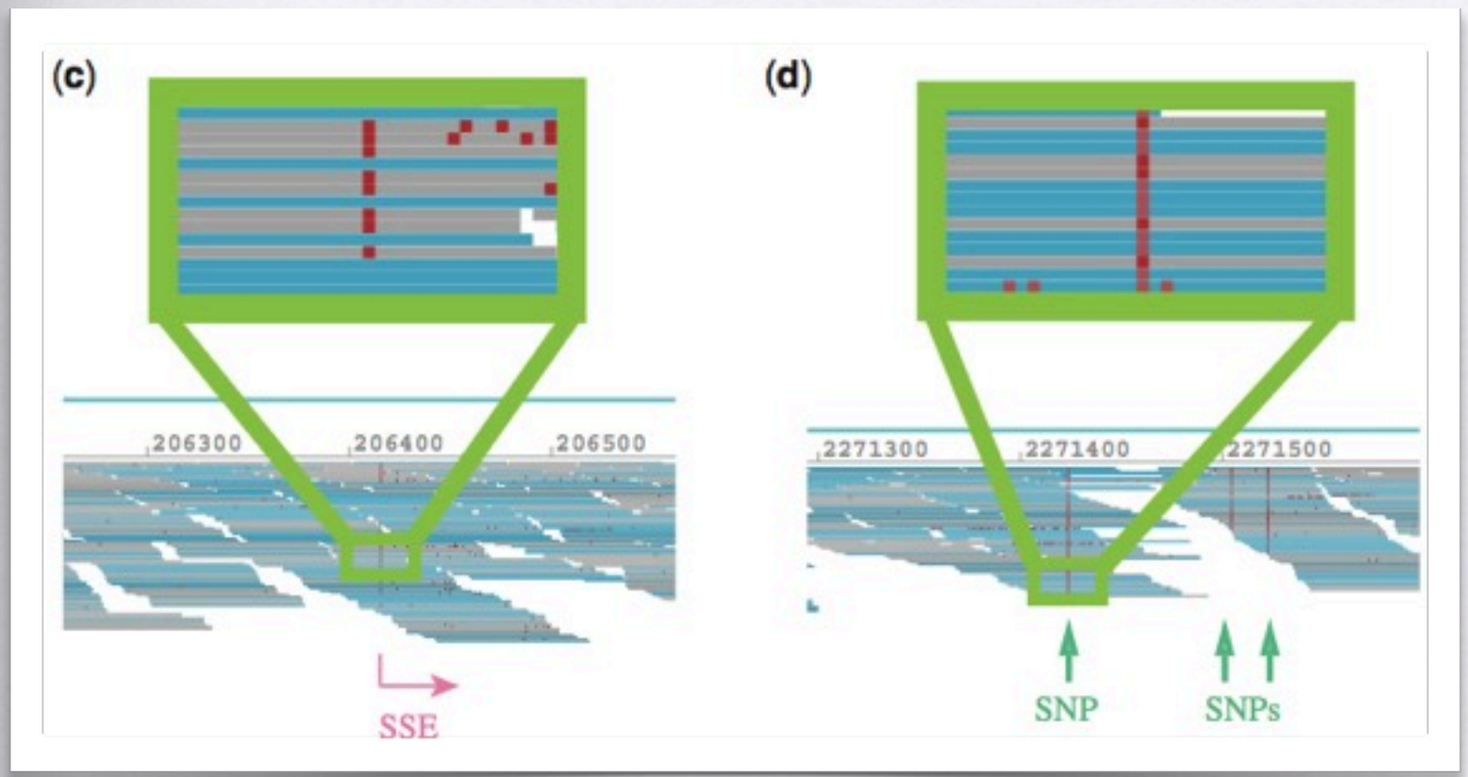


Figure 6. Schematic representation of the (**a**) inverted repeat and (**b**) enzyme preference for the SSE hypothetical mechanistic models. The gray numbers at the top indicate the cycle number and the numbers below indicate the relative population of each single-stranded DNA during the cycle. The colored bases and numbers below the drawings show the relative intensity of signals during that cycle. For instance, the second cycle of model (a) emits signals for C and G with an intensity of 73 and 27%, respectively.

Problems inherent to SSE

- Read depth coverage decreases in SSE regions (mismatches)
- SSE may cause false SNP calls
- Gaps in assembled sequences





THANKS!