Real-Time DNA Sequencing from Single Polymerase Molecules

John Eid,* Adrian Fehr,* Jeremy Gray,* Khai Luong,* John Lyle,* Geoff Otto,* Paul Peluso,* David Rank,* Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach,† Stephen Turner†

Originally published in *Science* Express on 20 November 2008
Science 2 January 2009:
Vol. 323. no. 5910, pp. 133 - 138

20th of January, 2009 Age Tats

- Nanophotonic structure – the zero-mode waveguide (ZMW).
  Ø 100 nm

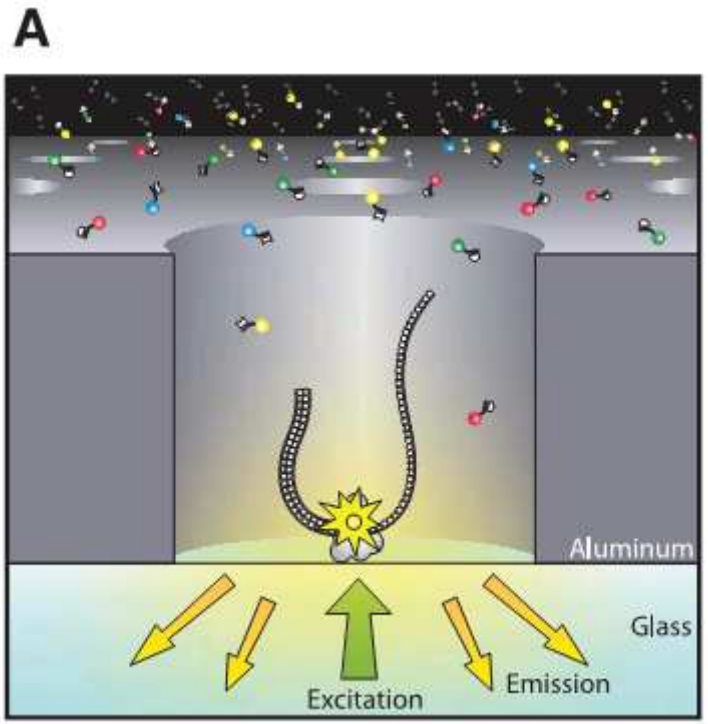
.

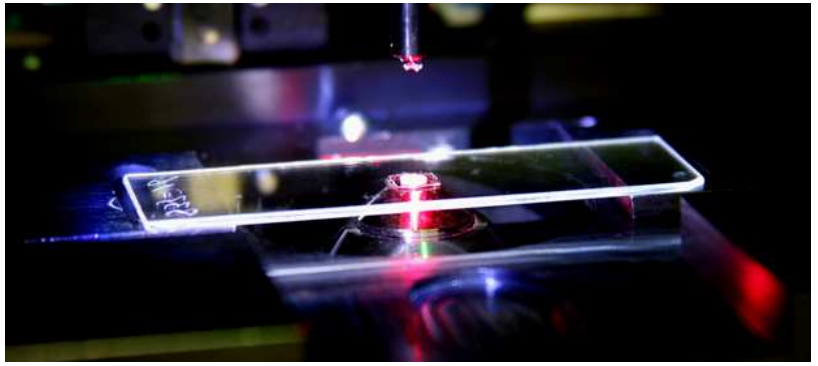
- Arranged in a rectangular array
  - 93 rows (spacing 1.3 μm)
  - 33 columns (spacing 4.0 μm)

- 37% of ZMWs – empty
- 37% of ZMWs – with one molecul
- 24% of ZMWs – with two or more molecules

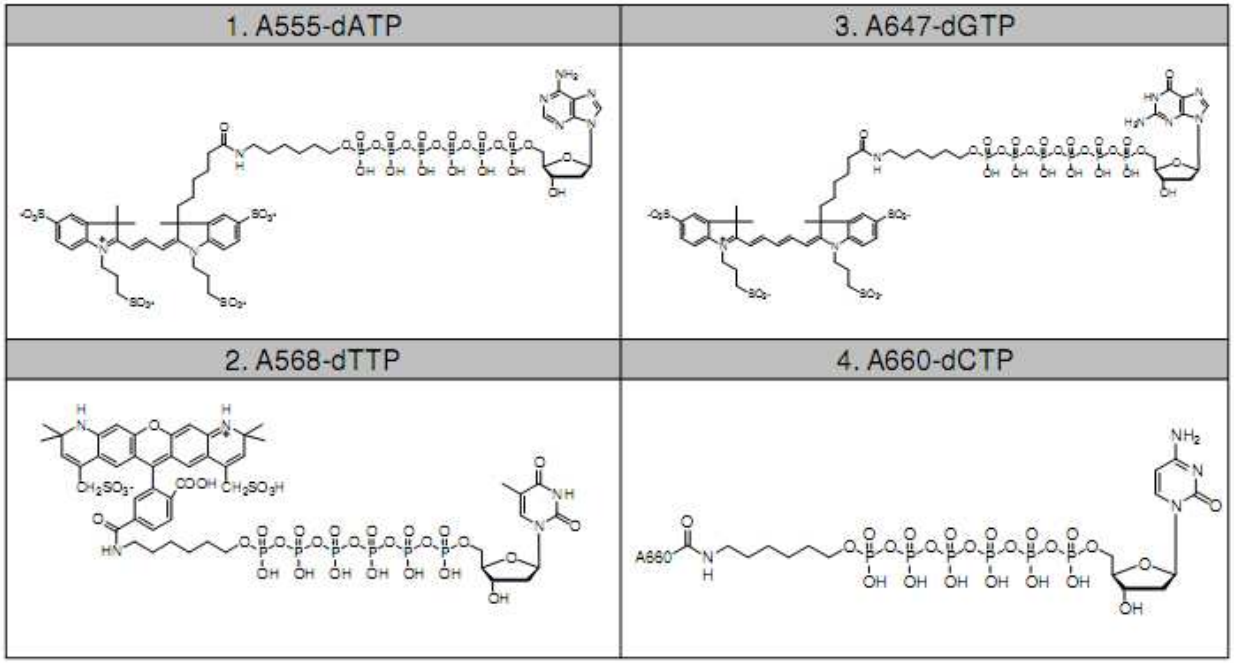
- Data were collected on a highly parallel confocal fluorescence detection instrument

- Bacteriophage Φ29 DNA polymerase

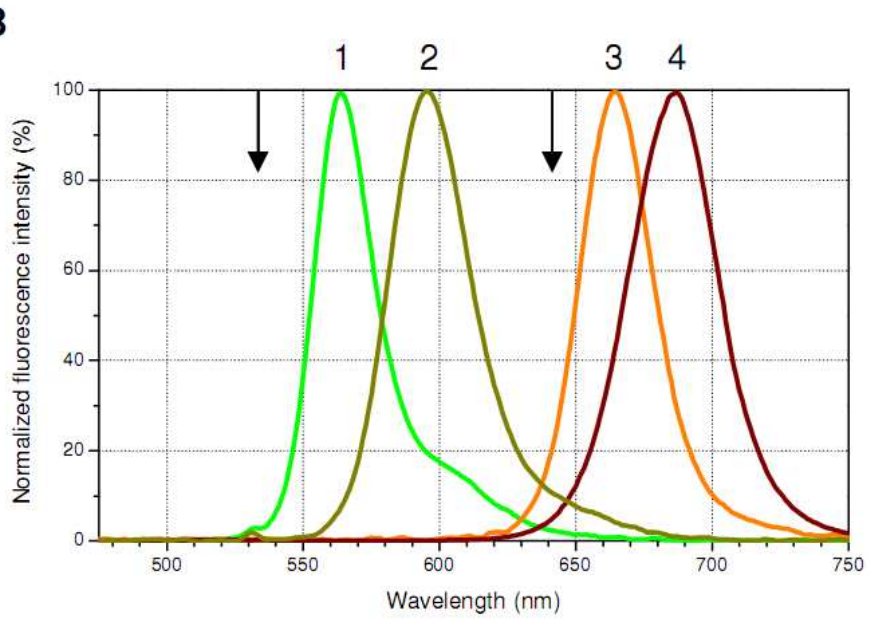




A


1. A555-dATP
3. A647-dGTP
2. A568-dTTP
4. A660-dCTP
A660

B


1
2
3
4
Normalized fluorescence intensity (%)
Wavelength (nm)
0
20
40
60
80
100
500
550
600
650
700
750

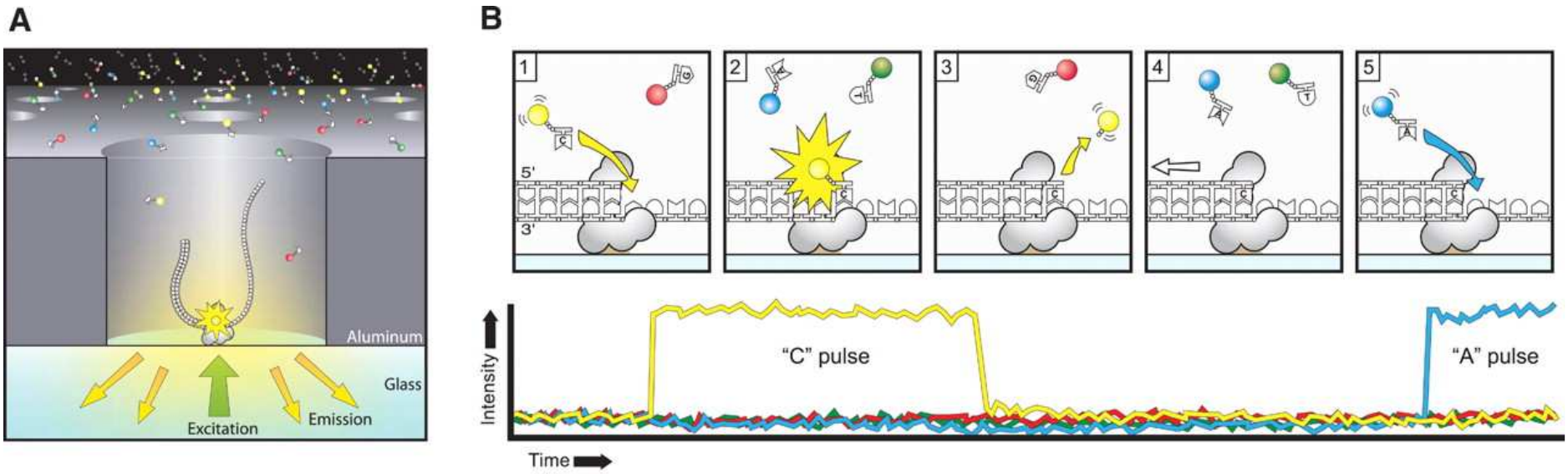
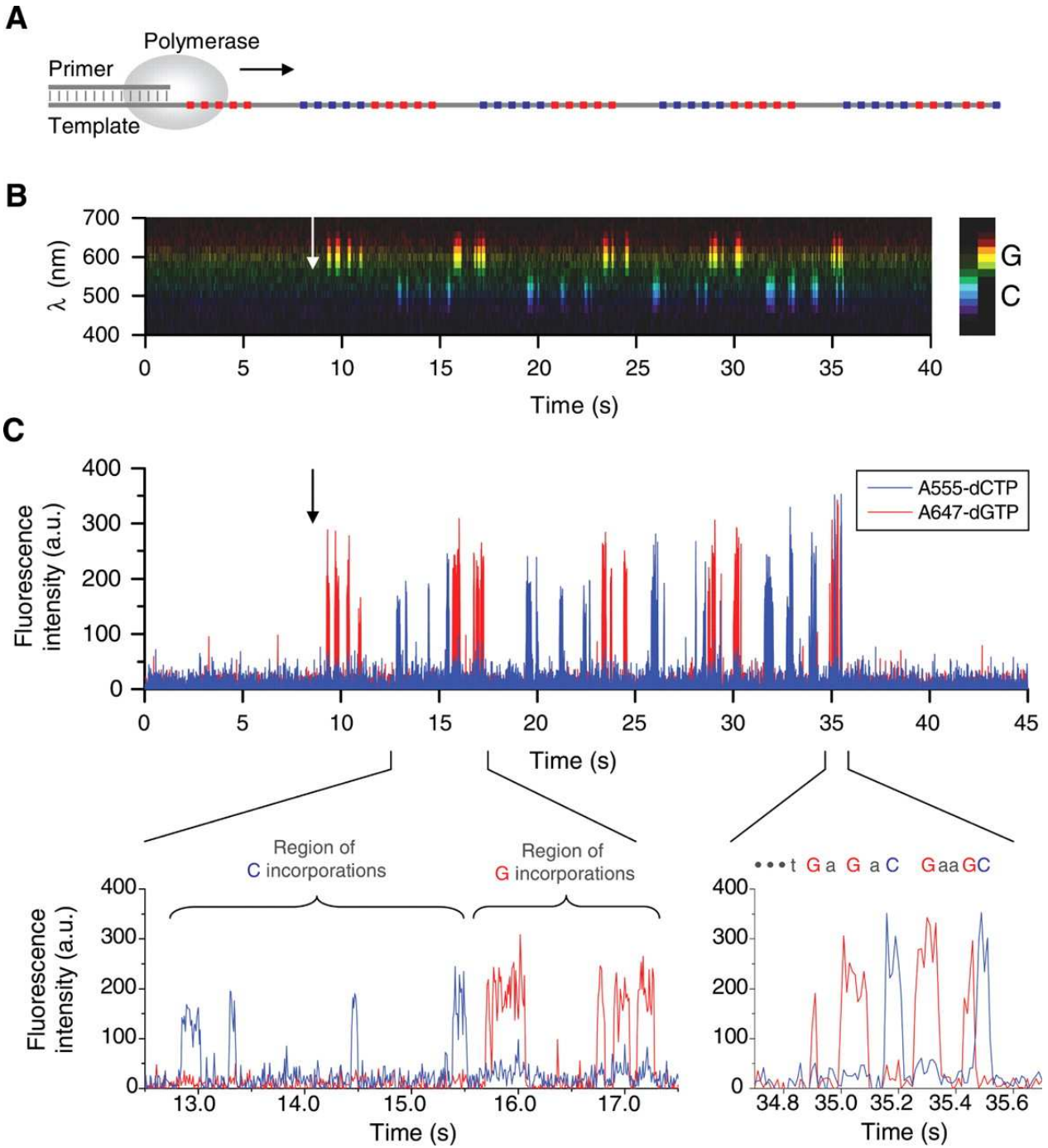


Fig. 1. Principle of single-molecule, real-time DNA sequencing. (**A**) Experimental geometry. A single molecule of DNA template-bound Φ29 DNA polymerase is immobilized at the bottom of a ZMW, which is illuminated from below by laser light. The ZMW nanostructure provides excitation confinement in the zeptoliter (10^{-21} liter) regime, enabling detection of individual phospholinked nucleotide substrates against the bulk solution background as they are incorporated into the DNA strand by the polymerase. (**B**) Schematic event sequence of the phospholinked dNTP incorporation cycle, with a corresponding expected time trace of detected fluorescence intensity from the ZMW. (1) A phospholinked nucleotide forms a cognate association with the template in the polymerase active site, (2) causing an elevation of the fluorescence output on the corresponding color channel. (3) Phosphodiester bond formation liberates the dye-linker-pyrophosphate product, which diffuses out of the ZMW, thus ending the fluorescence pulse. (4) The polymerase translocates to the next position, and (5) the next cognate nucleotide binds the active site beginning the subsequent pulse.


A
Polymerase
Primer
Template
B
λ (nm)
700
600
500
400
0
5
10
15
20
25
30
35
40
Time (s)
G
C
C
Fluorescence intensity (a.u.)
400
300
200
100
0
A555-dCTP
A647-dGTP
0
5
10
15
20
25
30
35
40
45
Time (s)
Region of C incorporations
Region of G incorporations
Fluorescence intensity (a.u.)
400
300
200
100
0
13.0
14.0
15.0
16.0
17.0
Time (s)
• • • t G a G a C G aa GC
400
300
200
100
0
34.8
35.0
35.2
35.4
35.6
Time (s)

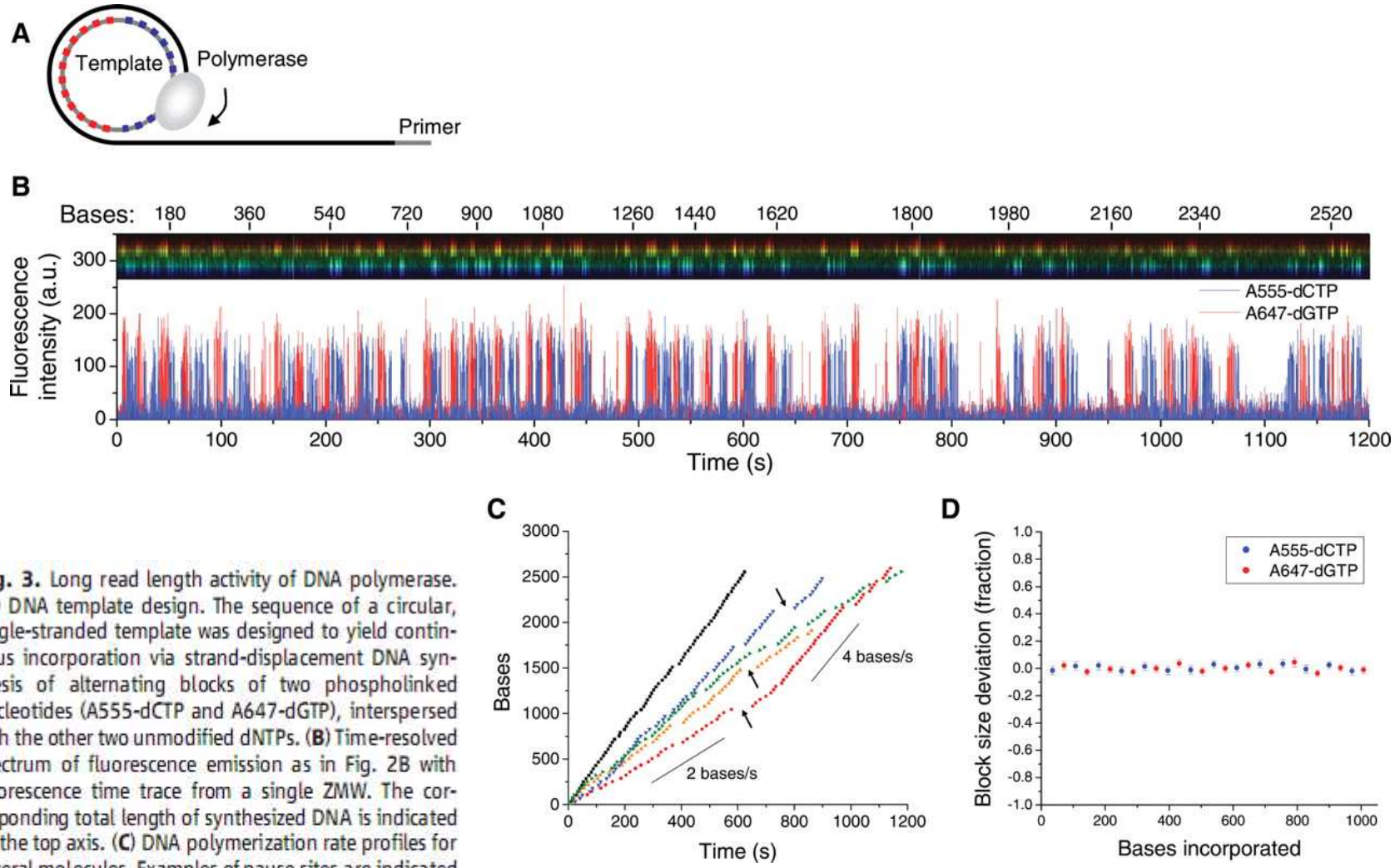


Fig. 3. Long read length activity of DNA polymerase. (**A**) DNA template design. The sequence of a circular, single-stranded template was designed to yield continuous incorporation via strand-displacement DNA synthesis of alternating blocks of two phospholinked nucleotides (A555-dCTP and A647-dGTP), interspersed with the other two unmodified dNTPs. (**B**) Time-resolved spectrum of fluorescence emission as in Fig. 2B with fluorescence time trace from a single ZMW. The corresponding total length of synthesized DNA is indicated by the top axis. (**C**) DNA polymerization rate profiles for several molecules. Examples of pause sites are indicated by arrows. The two lines indicate two persistent polymerization rates. (**D**) Error as a function of length of read for 14 rolling circle cycles (1008 total base incorporations; n = 186 reads). The fractional deviation from the average number of pulses per block (12 A555-dCTP and 12 A647-dGTP observed phospholinked dNTP pulses per cycle, respectively), mean ± SE, is plotted as a function of template position. The 95% confidence interval for the slope is −0.027 to +0.036 blocks per 1008 bases of incorporation.

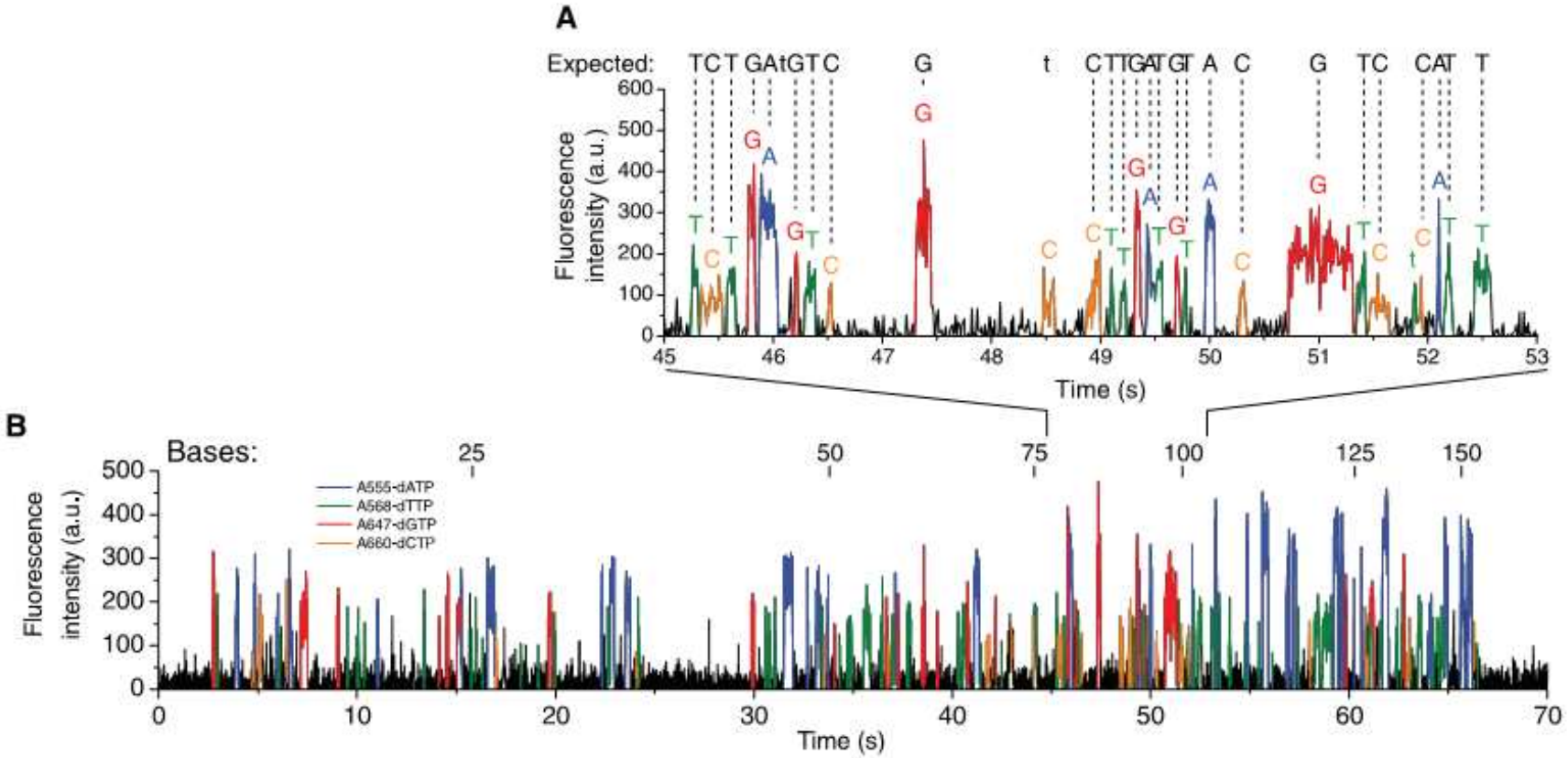


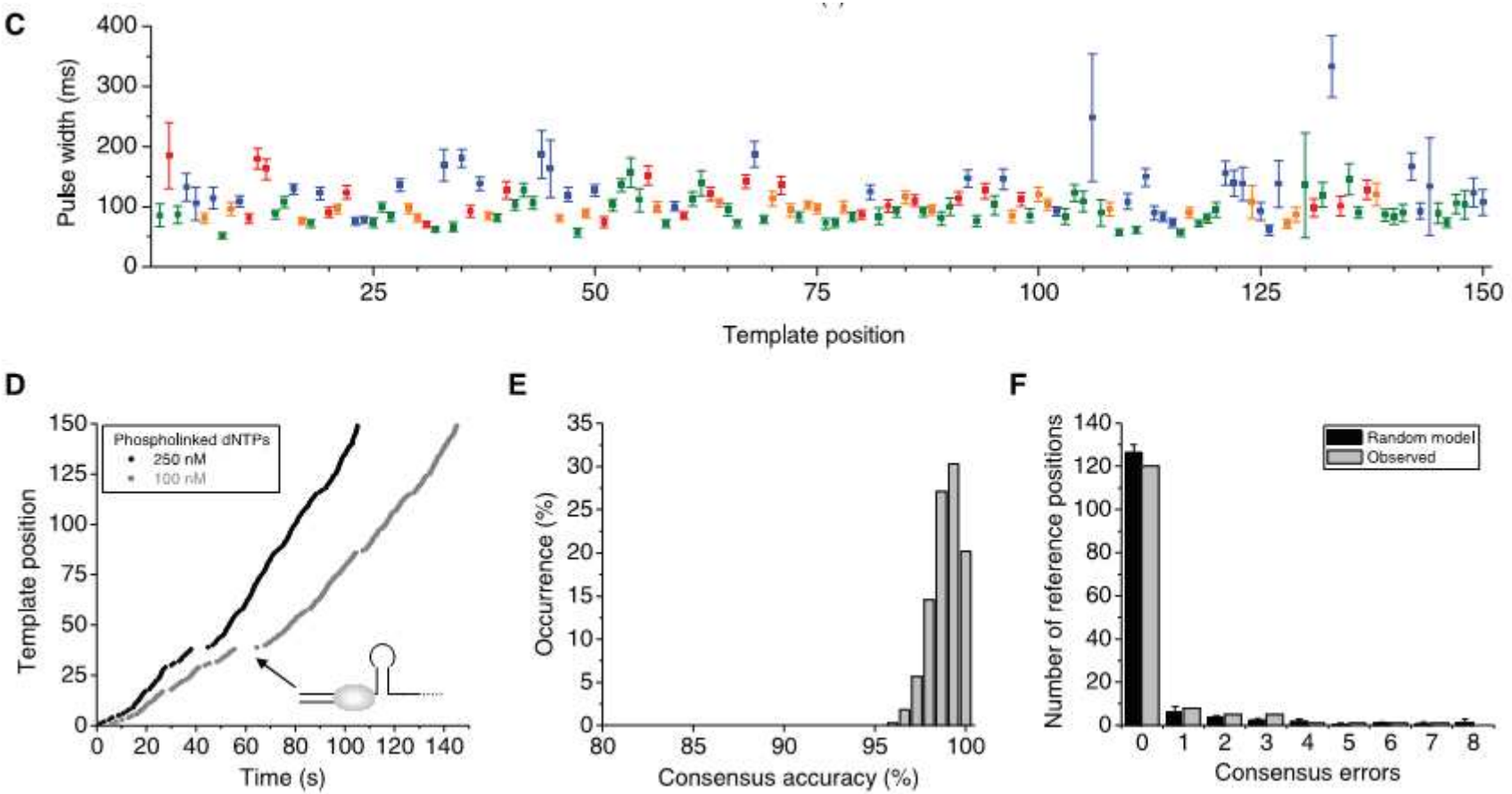
Fig. 4. Single-molecule, real-time, four-color DNA sequencing. (**A**) Total intensity output of all four dye-weighted channels, with pulses colored corresponding to the least-squares fitting decisions of the algorithm. This section of a fluorescence time trace shows 28 bases of incorporations and three errors. The expected template sequence is shown above, with dashed lines corresponding to matches; errors are in lowercase. (**B**) The entire read that proceeds through all 150 bases of the linear template. On average, ~63% of reads proceeded through the entire length of the DNA template.

Errors

- Deletions – undetected incorporations
  - Solutions: engineering the enzyme to reduce the fraction of short incorporation events; increasing fluorescence brightness; improving the efficiency of light collection

- Insertions – dissociation of a cognate nt from the active site before phosphodiester bond formation can occur -> erroneous duplication of a pulse
  - Solution: decreasing the dissociation rate before catalysis by decreasing the free-energy of the enzyme-substrate bound state

- Mismatches - spectral missassignments of the dyes
  - Solution: finding dye sets with larger spectral separations; increasing the brigthness of the dyes and collection efficiency of the instrument



(**C**) Average pulse width as a function of template position (extracted from $n = 449$ reads). (**D**) Cumulative interpulse duration plotted as a function of template position for two different phospholinked dNTP concentrations (250 nM, $n = 449$ reads; 100 nM, $n = 868$ reads). The arrow indicates a pause site observed for both conditions at position 40, corresponding to predicted secondary structure in the template at position 46 (fig. S7), taking into account the enzyme's footprint on the template (*42*). (**E**) Histogram of the sequence accuracy of 100 consensus sequences created by subsampling from 449 single-molecule reads to 15-fold average coverage. The median accuracy of the distribution is 99.3%. (**F**) Observed systematic bias compared with prediction from a random model free of sequence context bias. The error frequencies for observed (gray bars) and bias-free model data (black bars) are plotted in a histogram with the number of errors on the *x* axis and the number of different reference positions showing this many errors in 100 trials on the *y* axis. The random model is based on the observed error frequencies (table S3) (*26*).

Conclusions

- With just 15 molecules the median accuracy of 99.3%
- No detectable sequence bias
- Uniform error profile within reads

→ adequate for resequencing applications

Current limited experimental multiplex could be applied to sequencing small viral and bacterial genomes.

Each ZMW can produce sequence at a rate > 400 kb/day→

14000 functioning ZMWs→1-fold coverage of a diploid human genome per day.