Anal. Chem. **2011,** *83,* 509–515

A 265-Base DNA Sequencing Read by Capillary Electrophoresis with No Separation Matrix

Jennifer Coyne Albrecht,[†] Jennifer S. Lin,[‡] and Annelise E. Barron*[,†,‡]

Departments of Chemical Engineering and Bioengineering, Stanford University, 318 Campus Drive, W300B James H. Clark Center, Stanford, California 94305-5444, United States

Electrophoretic DNA sequencing without a polymer matrix is currently possible only with the use of some kind of "drag-tag" as a mobility modifier. In free-solution conjugate electrophoresis (FSCE), a drag-tag attached to each DNA fragment breaks linear charge-to-friction scaling, enabling size-based separation in aqueous buffer alone. Here we report a 265-base read for free-solution DNA sequencing by capillary electrophoresis using a random-coil protein drag-tag of unprecedented length and purity. We identified certain methods of protein expression and purification that allow the production of highly monodisperse drag-tags as long as 516 amino acids, which are almost charge neutral (+1 to +6) and yet highly water-soluble. Using a four-color LIF detector, 265 bases could be read in 30 min with a 267-amino acid drag-tag, on par with the average read of current next-gen sequencing systems. New types of multichannel systems that allow much higher throughput electrophoretic sequencing should be much more accessible in the absence of a requirement for viscous separation matrix.

Almost a decade after the draft sequence of the human genome was published, the number of DNA sequencing projects continues to grow exponentially.[1] The National Institutes of Health (NIH) initiative to decrease the cost of sequencing a human genome to $1K is essentially realized with recently published sequencing of $4.4K per genome (consumable costs only).[2] The advent of highly parallel, non-Sanger methods such as sequencing by hybridization and ligation[2,3] and by synthesis[4,5] has enabled this drastic decrease in cost. Third-generation sequencing methods under development are expected to further decrease the cost of whole genome sequencing. These methods include single molecule sequencing,[6] nanopores,[7] zero-mode wave guides,[8] and semiconductor pH sensing.[9] However, current next-generation methods require hours, if not days, to collect millions of bases of data per run,[10] and these instruments with highly complex optical detection are expensive and thus inaccessible to a majority of small research and medical laboratories.

The success of human organ transplantation directly correlates with the homology of 21 highly polymorphic genes located in the human leukocyte antigen (HLA) region of the genome.[11,12] Electrophoresis-based methods offer rapid, selective, and highly accurate sequencing of these few exons (<100) rather than the broad genome-wide results provided by ultrahigh-throughput next-generation technologies. Ideally, these 400−450 base long exons are sequenced entirely in one pass instead of with short reads (<100 bases) that must be assembled. Compared to next-generation technologies, electrophoresis offers the advantages of speed, ability to sequence through an entire exon at once, and only being parallelized[13] to the degree necessary for medical diagnostics such as HLA typing.[14] Traditional electrophoresis uses a polymeric sieving matrix to induce size-based separation of a ladder of Sanger fragments with average read lengths of 600−900 bases. This highly successful method was the backbone of the Human Genome Project and has been implemented onto microfluidic devices.[15] Miniaturization decreases analysis time and reduces sample volume, which is important for medical sequenc-

* Corresponding author. Phone: (650) 721-1151. Fax: (650) 723-9801. E-mail: aebarron@stanford.edu.

† Department of Chemical Engineering.

‡ Department of Bioengineering.

(1) Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. *Nucleic Acids Res.* **2010**, *38*, D46–51.

(2) Drmanac, R.; Sparks, A. B.; Callow, M. J.; Halpern, A. L.; Burns, N. L.; Kermani, B. G.; Carnevali, P.; Nazarenko, I.; Nilsen, G. B.; Yeung, G.; Dahl, F.; Fernandez, A.; Staker, B.; Pant, K. P.; Baccash, J.; Borcherding, A. P.; Brownley, A.; Cedeno, R.; Chen, L.; Chernikoff, D.; Cheung, A.; Chirita, R.; Curson, B.; Ebert, J. C.; Hacker, C. R.; et al. *Science* **2010**, *327*, 78–81.

(3) Shendure, J.; Porreca, G. J.; Reppas, N. B.; Lin, X. X.; McCutcheon, J. P.; Rosenbaum, A. M.; Wang, M. D.; Zhang, K.; Mitra, R. D.; Church, G. M. *Science* **2005**, *309*, 1728–1732.

(4) Margulies, M.; Egholm, M.; Altman, W. E.; Attiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z. T.; Dewell, S. B.; Du, L.; Fierro, J. M.; Gomes, X. V.; Godwin, B. C.; He, W.; Helgesen, S.; Ho, C. H.; Irzyk, G. P.; Jando, S. C.; Alenquer, M. L. I.; Jarvie, T. P.; Jirage, K. B.; Kim, J. B.; Knight, J. R.; et al. *Nature* **2005**, *437*, 376–380.

(5) Ju, J.; Kim, D. H.; Bi, L.; Meng, Q.; Bai, X.; Li, Z.; Li, X.; Marma, M. S.; Shi, S.; Wu, J.; Edwards, J. R.; Romu, A.; Turro, N. J. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 19635–19640.

(6) Pushkarev, D.; Neff, N. F.; Quake, S. R. *Nat. Biotechnol.* **2009**, *27*, 847–850.

(7) Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13770–13773.

(8) Levene, M. J.; Korlach, J.; Turner, S. W.; Foquet, M.; Craighead, H. G.; Webb, W. W. *Science* **2003**, *299*, 682–686.

(9) Karow, J. Ion Torrent Systems Presents $50,000 Electronic Sequencing at AGBT. In *Sequence*, March 2, 2010 (http://www.genomeweb.com/sequencing/ion-torrent-systems-presents-50000-electronic-sequencer-agbt).

(10) Voelkerding, K. V.; Dames, S. A.; Durtschi, J. D. *Clin. Chem.* **2009**, *55*, 641–658.

(11) Robinson, J.; Waller, M. J.; Parham, P.; Bodmer, J. G.; Marsh, S. G. E. *Nucleic Acids Res.* **2001**, *29*, 210–213.

(12) Field, S. F.; Nejentsev, S.; Walker, N. M.; Howson, J. M. M.; Godfrey, L. M.; Jolley, J. D.; Hardy, M. P. A.; Todd, J. A. *Diabetes* **2008**, *57*, 1753–1756.

(13) Paegel, B. M.; Emrich, C. A.; Weyemayer, G. J.; Scherer, J. R.; Mathies, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 574–579.

(14) Hert, D. G.; Fredlake, C. P.; Barron, A. E. *Electrophoresis* **2008**, *29*, 4618–4626.

(15) Fredlake, C. P.; Hert, D. G.; Kan, C. W.; Chiesl, T. N.; Root, B. E.; Forster, R. E.; Barron, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 476–481.

10.1021/ac102188p
Published on Web 12/23/2010

ing. However, loading high viscosity polymers into microfluidic devices requires high pressure and must be done carefully to avoid breaking chips; thus, it is not done in an automated fashion.

Free-solution conjugate electrophoresis (FSCE) has been under development for the past decade to eliminate the need for a sieving polymer when sizing DNA molecules (and is sometimes called "end-labeled free-solution electrophoresis," or ELFSE).[16−25] The mobility of DNA in free-solution electrophoresis with no polymer network is governed by the ratio of its charge to its friction. Because both scale linearly with length, mobility is rendered independent of length, and this "free-draining" polymer cannot be separated by size by electrophoresis in a buffer. (Free-solution electrophoresis in nanochannels can separate short pieces of DNA[26] but not with the resolution or read length necessary for medical diagnostic sequencing.) In FSCE, size-dependent mobility of DNA fragments is achieved by conjugating them to a mobility modifier ("drag-tag") with a different mobility than DNA. An ideal drag-tag is large enough to add significant friction to the DNA, water-soluble, completely monodisperse, almost charge neutral, can be uniquely and stably attached to the DNA, and interacts minimally with the microchannel walls. FSCE separations have been achieved with commercially available polymers[20] and proteins,[17,18,27] fluorophores,[28] surfactant micelles,[29,30] and chemically synthesized peptide mimics ("peptoids");[19,21,25] however, none of these drag-tags had sufficient purity and friction to achieve separations with single-base resolution of 400−450 bases of DNA. Initial proof-of-concept sequencing with the protein streptavidin yielded only 110 bases.[18]

To achieve longer sequencing reads by FSCE, a family of genetically engineered, highly repetitive "protein polymers" (expressed in *E. coli*) was developed. The repetitive protein has an amino acid block Gly-Ala-Gly-Thr-Gly-Ser-Ala, which is referred to herein as the "monomer" unit. The protein's monomer block was designed to encompass all the desired characteristics of a drag-tag: water solubility, charge neutrality, ease of conjugation using the single amine at the N-terminus, and random-coil structure to achieve high friction per unit. Charge neutrality is important because positively charged amino acids can adhere to negatively charged microchannel walls[18] and negatively charged amino acids can decrease the relative drag of the protein.[23] The stringent requirement for monodispersity ensures that only one peak is present for each length of DNA, which is critically important for identification and base-calling of sequencing fragments.

The first use of a protein polymer drag-tag for FSCE sequencing was published in 2008.[24] With 18 repeats of the "monomer" (127 amino acids total), the practically monodisperse protein enabled size-based separation of DNA in free-solution electrophoresis without a polymer network with a distinguishable sequence of 180 bases. Mutations in the protein introduced by *E. coli* changed 2 of the 18 uncharged serine residues to positively charged arginines, which increased the friction of the drag-tag but did not cause noticeable interactions with the capillary walls. Separations were diffusion-limited; unlike sieving polymer-based separations, increased electric field increased the speed of the separation without inducing band broadening. On the basis of this data, the only limitation to longer sequencing read lengths (400−450 bases for an entire exon) was lack of a larger drag-tag with higher friction. Initial attempts to produce larger proteins based on the same repetitive monomer were unsuccessful due to heterogeneity. Modifications to the expression vector and purification system were necessary to achieve larger protein polymer drag-tags suitable for DNA sequencing.[31]

In this paper, we present four large protein polymer drag-tags, with 27−72 repeats of the monomer unit, which are sufficiently purified for FSCE separations. The 27mer and 36mer proteins (204 and 267-aa total length) are used for sequencing by free-solution electrophoresis with no sieving polymer. The 36mer drag-tag separated ∼265 bases of sequencing fragments, which is almost a 50% increase in read length over the 18mer drag-tag. The 54mer and 72mer drag-tags (390- and 516-aa total length), although pure enough for good FSCE separations, were unusable for sequencing by our current approaches due to an apparent inhibition by these proteins of the Sanger cycle sequencing reaction when appended to the 5′ end of the sequencing primer. On the basis of the 27 and 36mer results, the 72mer drag-tag theoretically will separate the minimum of 400 bases of DNA with high resolution necessary to sequence through an exon and will put FSCE on par with read lengths from next-generation sequencing instruments. Since the protein monomer was based on the successful 18mer, where 1 of every 9 monomers contained a positively charged arginine, these proteins have 3−8 positively charged amino acids. In addition to describing the FSCE sequencing results, this paper will explore the use of these longer protein polymer drag-tags with increased charge for sequencing by FSCE and the impact of the charges on peak separations.

MATERIALS AND METHODS

Drag-Tag Production and Purification. Using previously described methods,[31,32] highly repetitive, genetically engineered "protein polymers" were produced in *E. coli*. Four lengths of

(16) Mayer, P.; Slater, G. W.; Drouin, G. *Anal. Chem.* **1994**, *66*, 1777–1780.
(17) Heller, C.; Slater, G. W.; Mayer, P.; Dovichi, N.; Pinto, D.; Viovy, J. L.; Drouin, G. *J. Chromatogr. A* **1998**, *806*, 113–121.
(18) Ren, H.; Karger, A. E.; Oaks, F.; Menchen, S.; Slater, G. W.; Drouin, G. *Electrophoresis* **1999**, *20*, 2501–2509.
(19) Vreeland, W. N.; Barron, A. E. *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **2000**, *41*, 1018–1019.
(20) Vreeland, W. N.; Desruisseaux, C.; Karger, A. E.; Drouin, G.; Slater, G. W.; Barron, A. E. *Anal. Chem.* **2001**, *73*, 1795–1803.
(21) Vreeland, W. N.; Slater, G. W.; Barron, A. E. *Bioconjugate Chem.* **2002**, *13*, 663–670.
(22) Meagher, R. J.; Won, J. I.; McCormick, L. C.; Nedelcu, S.; Bertrand, M. M.; Bertram, J. L.; Drouin, G.; Barron, A. E.; Slater, G. W. *Electrophoresis* **2005**, *26*, 331–350.
(23) Won, J. I.; Meagher, R. J.; Barron, A. E. *Electrophoresis* **2005**, *26*, 2138–2148.
(24) Meagher, R. J.; Won, J. I.; Coyne, J. A.; Lin, J.; Barron, A. E. *Anal. Chem.* **2008**, *80*, 2842–2848.
(25) Haynes, R. D.; Meagher, R. J.; Won, J. I.; Bogdan, F. M.; Barron, A. E. *Bioconjugate Chem.* **2005**, *16*, 929–938.
(26) Pennathur, S.; Baldessari, F.; Kattah, M. G.; Steinman, J. B.; Utz, P. J.; Santiago, J. G. *Anal. Chem.* **2007**, *79*, 8316–8322.
(27) Lau, H. W.; Archer, L. A. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2010**, *81*, 031918.
(28) Sudor, J.; Novotny, M. V. *Anal. Chem.* **1995**, *67*, 4205–4209.
(29) Grosser, S. T.; Savard, J. M.; Schneider, J. W. *Anal. Chem.* **2007**, *79*, 9513–9519.
(30) Savard, J. M.; Grosser, S. T.; Schneider, J. W. *Electrophoresis* **2008**, *29*, 2779–2789.
(31) Lin, J. S.; Albrecht, J. C.; Wang, X.; Meagher, R. J.; Barron, A. E. *Biomacromolecules* **2010**, submitted.
(32) Won, J. I.; Barron, A. E. *Macromolecules* **2002**, *35*, 8281–8287.

protein were produced based on the repeating amino acid sequence Gly-Ala-Gly-Thr-Gly-Ser-Ala (with 27, 36, 54, and 72 repeating units); 1 in every 9 repeating units contained arginine in place of serine due to a mutation introduced by *E. coli*.[24,31] A "controlled cloning" method was used to assemble the DNA sequence of the desired protein[32] with a T7 promoter sequence (MASMTGGQQMG) at the N-terminus for enhanced expression and $IEGRH_8$ at the C-terminus for purification. The proteins were expressed in *E. coli*, and the full-length protein was recovered from cell lysate by affinity chromatography with Talon cobalt-chelated resin (Clontech, Mountain View, CA). After purification by RP-HPLC and removal of the histidine affinity tag by endoproteinase GluC, the protein retained the negatively charged glutamic acid residue at its C-terminus. The protein was dried on a lyophilizer and stored at −20 °C until further use.

Conjugation of Drag-Tag and Sequencing Sample Preparation. To conjugate protein drag-tags to DNA (sequencing primer: 5′-X_1-GTT TTC CCA GTC ACG AC; 30-nt: 5′-X_1-CC-X_2-TTT AGG GTT TTC CCA GTC ACG ACG TTG, where X_1 = 5′-C6 thiol linker, X_2 = dT-fluorescein; Integrated DNA Technologies, Coralville, IA), they were activated with the heterobifunctional linker molecule sulfo-SMCC (sulfosuccinimidyl 4-*N*-maleimidomethyl cyclohexane-1-carboxylate, Thermo Fisher Scientific, Waltham, MA) using a previously described protocol.[24,33] In short, the protein was mixed with a 10:1 molar excess of sulfo-SMCC, vortexed for 1 h, and lyophilized after using a CentriSep gel filtration column (Princeton Separations, Adelphia, NJ) to remove excess sulfo-SMCC. The thiol-terminated ssDNA oligomer was reduced with a 20:1 molar excess of TCEP (tris(2-carboxyethyl)phosphine, Thermo) at 40 °C for 100 min, desalted, and separated from excess TCEP with a CentriSep column and then incubated with a 100:1 molar excess of the activated drag-tag at room temperature for 4−18 h.

To test the conjugation of the drag-tag to the sequencing primer, a single-base extension (SBE) assay was performed. A 2.2 pmol amount of DNA−drag-tag conjugate, 62.5 ng of M13mp18 ssDNA template (New England Biolabs, Ipswich, MA), 5.0 μL of SNaPshot Multiplex mix (Applied Biosystems, Foster City, CA), and water were mixed to a total volume of 10 μL. The reaction was heated at 96 °C for 1 min then cycled 25 times: 96 °C for 10 s, 50 °C for 5 s, and 60 °C for 30 s (Eppendorf Mastercycler Gradient). The sample was purified with a CentriSep column, denatured at 95 °C for 2 min, and snap-cooled on ice for 5−10 min. To create the sequencing sample, the following was mixed: 8.4 pmol of sequencing primer plus drag-tag, 0.16 μg of M13mp18 ssDNA template, 8.0 μL of BigDye terminator v1.1 cycle sequencing mix (ABI), and water to a total volume of 20 μL. After incubation at 96 °C for 1 min, the sequencing reaction was cycled 36 times (96 °C for 10 s, 50 °C for 5 s, 60 °C for 30 s to 2 min). The sample was purified, denatured, and snap-cooled as described above.

Capillary Electrophoresis. Separations of drag-tags plus ssDNA oligomers or DNA sequencing fragments were performed using an Applied Biosystems Prism 3100 Genetic Analyzer with four-color LIF detection. The 16-capillary array of bare fused-silica capillaries has an inlet-to-detector length of 36 cm (total length 47 cm) and 50 μm ID. Electrophoresis was performed in 1 X TTE buffer (89 mM Tris, 89 mM TAPS, 2 mM EDTA) plus 7 M urea and a 1:200 dilution of POP-6 ("Performance-Optimized Polymer", ABI) for dynamic wall-coating.[24,25,33,34] The drag-tagged samples were introduced into the capillary array by electrokinetic injection at 22 V/cm for 20 s, and the separation was carried out at 55 °C with an electric field strength of 62−312 V/cm (3−15 kV applied voltage). Fresh buffer was flushed into the array between each run, and reservoirs were refilled every 1−5 runs.

RESULTS AND DISCUSSION

Drag-tags were evaluated for sequencing by free-solution electrophoresis with no polymer network when conjugated to an ssDNA oligomer of known length. Successful conjugates migrate slower than free-draining unconjugated DNA. The added friction of the drag-tag is measured experimentally. FSCE theory describes the mobility of the DNA−drag-tag conjugate with the following equation, assuming that that drag-tag is uncharged and in a random-coil conformation under the experimental conditions:[22]

$$\mu = \mu_0\left(\frac{M_c}{M_c + \alpha_1 M_u}\right) \quad (1)$$

where μ is the mobility of the DNA−drag-tag conjugate, μ_0 is the free-solution mobility of unconjugated DNA, M_c is the number of charged monomers in the DNA (nucleotides), M_u is the number of neutral monomers in the drag-tag, and α_1 is the friction coefficient of each uncharged drag-tag monomer. The value of $\alpha = \alpha_1 M_u$ is the overall drag from the drag-tag. The "effective" α value (assuming the drag-tags are uncharged even though they contain up to 8 positively charged amino acids) was determined experimentally from the mobilities of the unconjugated DNA and the DNA−drag-tag conjugate with units equivalent to the number of ssDNA bases that impart the same amount of friction. Therefore, the higher the α value, the longer the achievable sequencing read length.

Drag-Tag Evaluation. New methods were recently developed to produce monodisperse protein polymers of greatly increased length.[31] These proteins are designed with an elongated, random-coil structure instead of a globular, compact structure to achieve the most friction per unit length. Protein polymers were produced with 27, 36, 54, and 72 repeating units (204, 267, 390, and 516 amino acids in length, respectively, with 1 of every 9 serine residues mutated to an arginine)[24,31] and purified to almost full monodispersity. After conjugation to DNA, the resulting proteins have a net charge of +1, +2, +4, and +6 (for the 27-, 36-, 54-, and 72mer drag-tags, with one negative charge in the affinity tag). The sequencing ability of the 18mer protein was not adversely affected by 2 additional arginines (net +1 with a different expression system).[24] The impact of the increase in charge on FSCE sequencing with these longer proteins will be examined.

The electropherogram in Figure 1A shows the successful conjugation of the 27- and 36mer proteins to the sequencing primer (18-nt long). Both proteins have large peaks for the DNA−drag-tag conjugates with minimal extra peaks in the

(33) Meagher, R. J.; Coyne, J. A.; Hestekin, C. N.; Chiesl, T. N.; Haynes, R. D.; Won, J. I.; Barron, A. E. *Anal. Chem.* **2007**, *79*, 1848–1854.

(34) Meagher, R. J.; McCormick, L. C.; Haynes, R. D.; Won, J. I.; Lin, J. S.; Slater, G. W.; Barron, A. E. *Electrophoresis* **2006**, *27*, 1702–1712.

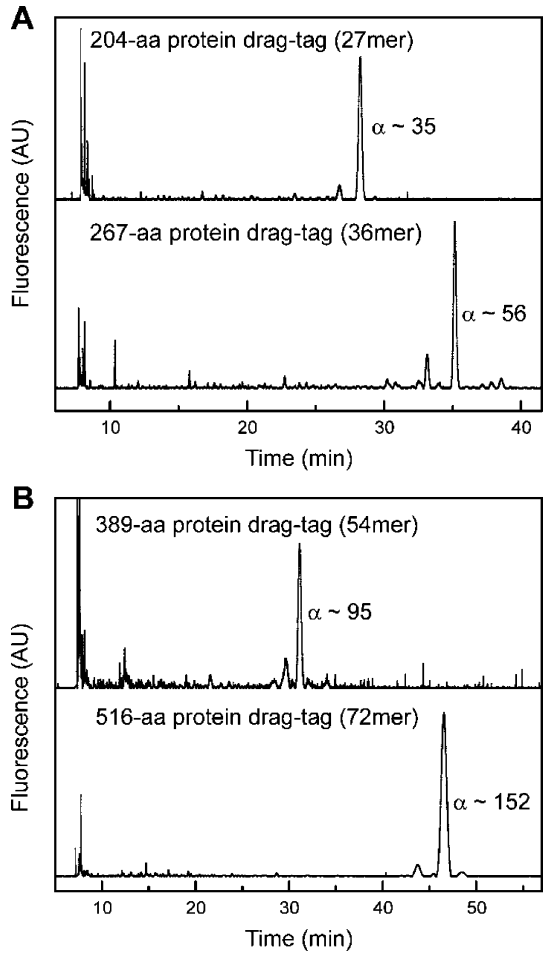

Figure 1. Electrophoretic analysis of the purity of increasing lengths of protein drag-tags; the cluster of peaks at 7 min is "free" unconjugated DNA, and the second peak is the DNA−drag-tag conjugate. Measured α values are noted. (A) 27mer (204-aa) and 36mer (267-aa) proteins conjugated to M13 sequencing primer, post-single-base extension reaction (18-nt oligomer). (B) 54mer (390-aa) and 72mer (516-aa) proteins conjugated to 30-nt oligomer. Electrophoresis is performed on ABI 3100 (36-cm capillary) with electrokinetic injection at 22 V/cm for 20 s, separation at 312 V/cm, 55 °C, in 1 X TTE buffer with 7 M urea and 1:200 dilution of POP-6 as a wall coating.

baseline. The 27mer protein has an α of 35 while the 36mer protein has an α of 56 which is more than twice the α of the 18mer drag-tag used to sequence 180 bases of DNA.[24] The small extra peaks are similar to those seen in the baseline of the 18mer[24] and are not expected to decrease the efficiency of sequencing since the 18mer was not impacted negatively. Figure 1B shows the 54mer and 72mer proteins conjugated to a fluorescently labeled 30-nt oligomer with α values of 95 and 152, respectively.

Sequencing with Longer Protein Polymer Drag-Tags. Sequencing fragments were generated using the sequencing primer conjugated to the 27mer and 36mer drag-tags. The sequencing fragments were successfully separated by free-solution electrophoresis with no entangled polymer network present. Representative electropherograms are shown in Figure 2 (36mer drag-tag) and Supporting Information Figure S-1 (27mer drag-tag), with separations at 312 V/cm. The smallest fragment (18 bases) elutes last while the largest fragments migrate fastest; the sequence is read "backwards," starting at the right side of the bottom panel of the figures. The sequencing electropherograms are essentially "raw" data; the only corrections made were spectral deconvolution of the dyes (automatically performed by the ABI 3100) and baseline subtraction. No corrections have been made to normalize for peak height or mobility shifts induced by different dyes. The sequence obtained with the 36mer drag-tag was determined to $M = 170$ bases before repeated peaks become unresolved or peaks begin to overlap or become out of order due to different mobility shifts of the four dye molecules. Using the known sequence of the template for alignment, sequencing peaks separated with the 36mer drag-tag can be read to $M = 265$ bases. (Sequencing peaks separated with the 27mer protein are distinguishable to $M = 210$ bases.) The 36mer drag-tag is twice the length of the 18mer, has an α more than double, and enables approximately a 47% increase in read length.[24] This is the longest sequencing read ever recorded by FSCE separations, and longer drag-tags should theoretically give even longer reads.

The protein polymer drag-tags were incorporated into the traditional Sanger reaction with ease, which provides a notable advantage. The primers were conjugated to the drag-tags and included in the reaction without modification to the standard cycling protocol. While the previous study used the SNaPshot kit, this study used the BigDye kit (both ABI), demonstrating that the method is kit-independent. Both yield sequencing peaks with no sign of degradation from the presence of the drag-tag.[24] This advantage appears to be limited to proteins <390-aa, as neither the SBE nor the sequencing reaction proceeded with the two largest drag-tags conjugated to the primer (54mer with 390-aa, 72mer with 516-aa). The presence of these large proteins appended to the 5′ end of the primer inhibited the Sanger reaction, likely from some type of steric hindrance (the drag-tag could have blocked the hybridization of primer to template, or binding of polymerase to primer−template hybrid, or a combination of both). Control Sanger extension reactions were performed with standard primers where the appropriate amount of either the 54mer or 72mer drag-tags was spiked into the amplification reaction mixtures. When separated by a sieving polymer (POP-6), sequencing fragments generated in the presence of the large protein polymers were the same as the control reaction with no protein spiked into it, showing that the protein drag-tags only inhibit the sequencing reaction if they are attached to the primer (data not shown). Thus, a post-PCR conjugation method must be developed to take full advantage of these large protein drag-tags to sequence >265 bases of DNA.

Sequencing Peak Analysis: Band Broadening. When sequencing peaks obtained with the 27mer and 36mer protein drag-tags are compared to those with the 18mer protein drag-tag, band broadening is evident. The peak width (fwhm, w) was determined from the raw data, normalized by the speed of each fragment, and is shown versus DNA size in Figure 3. Peak width increases for any length of DNA as the drag-tag size increases. For each drag-tag, peak width decreases as length of DNA increases until it appears to reach somewhat of a horizontal asymptote.

To investigate band broadening, the plate height H was examined. The band broadening sources present in free-solution electrophoretic separations can be described with this Van Deemter-like equation, assuming negligible Joule heating:[18,24]

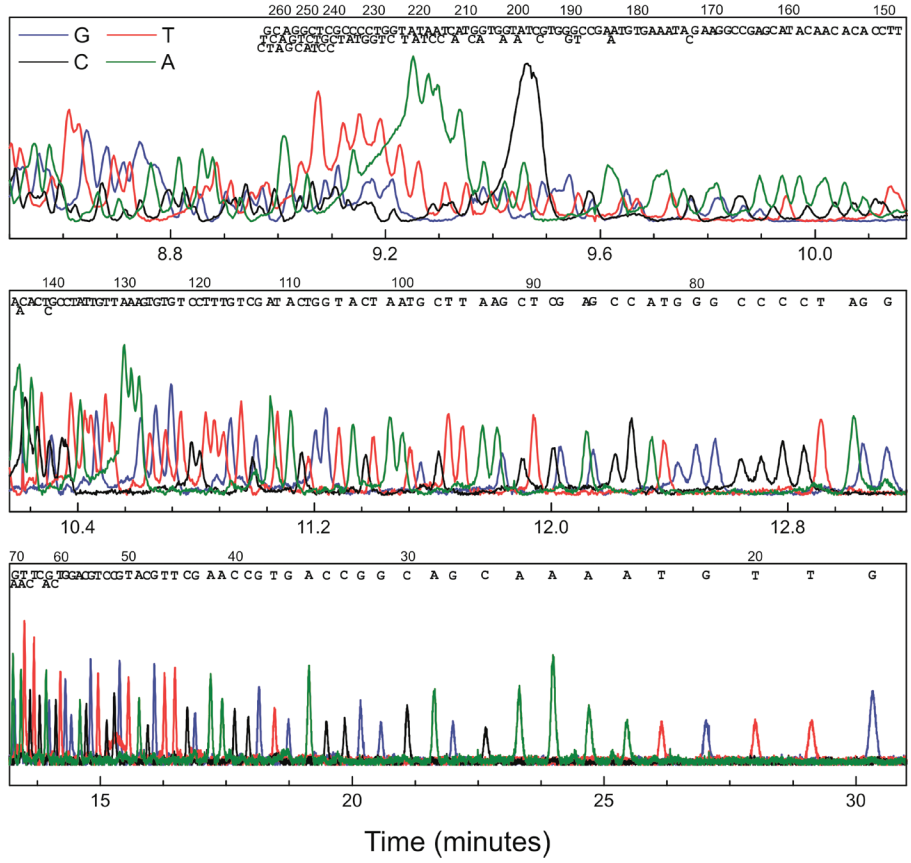

Figure 2. Four-color sequencing electropherogram with 36mer drag-tag (267-aa); 265 bases are resolved by electrophoresis without a sieving polymer under the same conditions as in Figure 1. M13mp18 template is "read" backward, starting at the right of the bottom panel.

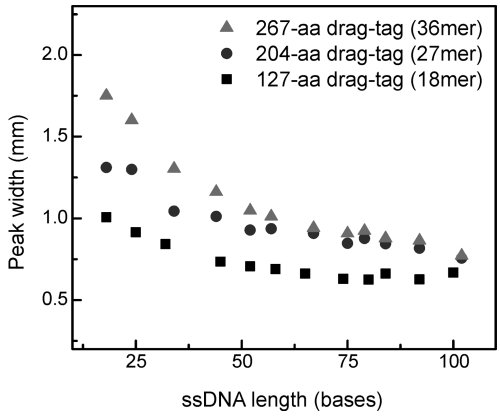

Figure 3. Peak width (fwhm) is plotted versus length of DNA sequencing fragment. At any length of DNA, peak width increases as drag-tag size increases.

$$H = \frac{A}{L} + \frac{2D}{u} + Wu + BL \qquad (2)$$

where the four possible sources of band broadening are (i) injection plug width, (ii) thermal diffusion, (iii) analyte−wall interactions, and (iv) drag-tag polydispersity; A, W, and B are constants related to i, iii, and iv, respectively, u is the electrophoretic velocity ($u = \mu E$), and L is the separation length (inlet to detector). The plate height H is determined from the raw data with this equation:

$$H = \frac{\sigma_x^2}{L} = \frac{w^2 u^2}{8L \times \ln(2)} \qquad (3)$$

where σ_x^2 is the spatial peak variance and is related to the temporal peak variance σ^2 ($\sigma^2 = \sigma_x^2/u^2$), which is related to temporal peak width w [$w^2 = \sigma^2 8\ln(2)$]. By varying u and L, two sets of experiments can examine all four possible causes of band broadening. However, the CE instrument (ABI 3100) is limited to 4 lengths of arrays, and previous work showed that no correlations about injection width and drag-tag polydispersity could be made with only four data points.[24]

To determine the impact of thermal diffusion and analyte−wall interactions on band broadening during FSCE separations, sequencing fragments with both the 27mer and 36mer drag-tags were separated at 9 electric field strengths ranging from E = 62−312 V/cm (applied voltage of 3−15 kV, increasing by increments of 1.5 kV). The plate height H was determined for two fragments (61-bp "C" and 104-bp "A"-terminated fragments), which are both distinct with no interference from neighboring peaks.[18,24] A graph of H vs u^{-1} for both drag-tags is shown in Figure 4 (27mer in 4A and 36mer in 4B). For a diffusion-limited separation, the H values should decrease linearly with increased u (decreased $1/u$).[24] For both the drag-tags, H trends along a straight line with the exception of the fragments separated at 312 V/cm (highest electric field possible with the instrument). The slope of this line can be used to estimate the diffusion coefficient; we found D of approximately 7.2×10^{-7} cm^2/s for separations with the 27mer drag-tag (slope $\sim 1.4 \times 10^{-4}$ mm^2/s) and D of approximately

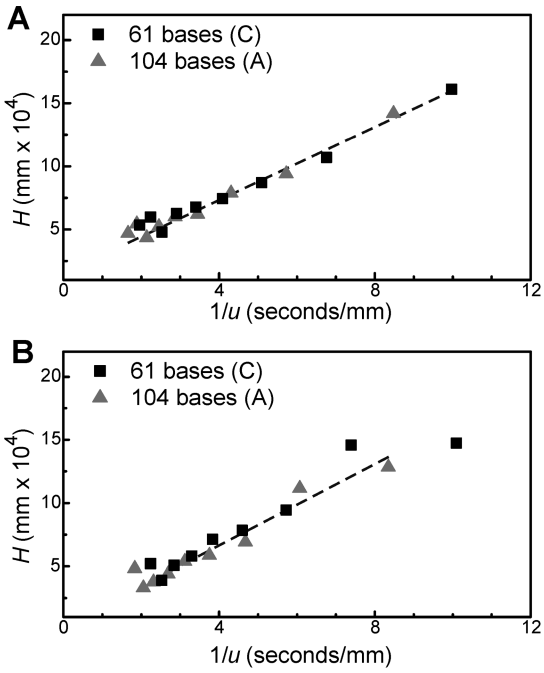

Figure 4. Separation peak analyses of a series of sequencing separations performed at 9 electric field strengths ($E = 62-312$ V/cm, applied voltage of 3−15 kV). Plate height H is plotted versus reciprocal speed for the (A) 27mer and (B) 36mer protein drag-tag. H was measured for two lengths of DNA, 61 bases (C-terminated) and 104 bases (A-terminated). Dashed line is the linear fit of H vs u^{-1} data for all but the highest E.

8.1×10^{-7} cm^2/s with the 36mer drag-tag (slope $\sim 1.6 \times 10^{-4}$ mm^2/s), both of which agree well with previous diffusion coefficient measurements of DNA in free-solution electrophoresis.[35] In all experiments, the bare fused-silica capillary walls were coated dynamically with POP-6 to suppress electroosmotic flow (EOF) and minimize interactions between the protein and the wall. The increase in H at high u (where $E = 312$ V/cm) indicates that the band broadening is no longer simply diffusion-limited and that analyte-wall interactions may be contributing to the peak broadness. This velocity series is limited by the maximum voltage of the CE instrument, however, and more data points at higher electric field strengths would be necessary to confirm wall effects as the major cause of band broadening.

While plate height H is an informative measure of separation efficiency and the causes of band broadening, it does not predict sequencing read length, which will be an important metric in determining how to balance the use of increased positive charges for added friction with increased peak width. Separation factor S (sometimes called "resolution length"[36] or "resolution"[24]) is a metric to evaluate FSCE sequencing separations with the same drag-tag and varied electrophoresis conditions:[18,24]

$$S = \frac{\bar{w}}{|\Delta t/\Delta M|} = \frac{1}{2}(w_1 + w_2)\left(\frac{M_1 - M_2}{t_2 - t_1}\right) \qquad (4)$$

(35) Nkodo, A. E.; Garnier, J. M.; Tinland, B.; Ren, H. J.; Desruisseaux, C.; McCormick, L. C.; Drouin, G.; Slater, G. W. *Electrophoresis* **2001**, *22*, 2424–2432.

(36) Heller, C. *Electrophoresis* **1999**, *20*, 1978–1986.

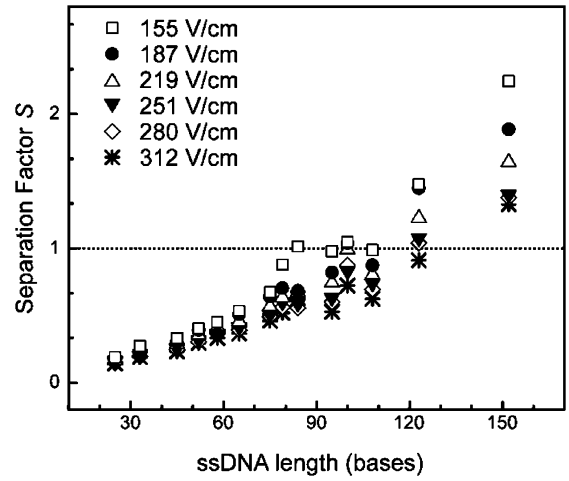

Figure 5. Separation factor S versus DNA length from sequencing data with the 36mer drag-tag are compared at varied E (155−312 V/cm, fragments are considered well-resolved if $S \leq 1$). The highest field strength gave the lowest S and is predicted to give the longest sequencing read.

where t_1 and t_2 are peak migration times, and M_1 and M_2 are the number of nucleotides in the respective ssDNA fragments, where $M_1 > M_2$. S is a discrete approximation of resolution which accounts for both peak width and spacing. An S value ≤ 1indicates a well-resolved pair of peaks where the distance between them is greater than the average of their widths. (S is opposite the traditional resolution R used to assess matrix-based sequencing separations, where a pair of peaks with $R \geq 0.5$ is considered well-resolved.)

Separation factor S for 14 pairs of peaks was determined for separations with the 36mer drag-tag at 6 electric field strengths ($M_1 = 26$, 34, 46, 53, 59, 66, 76, 80, 85, 96, 101, 109, 124, and 153 nucleotides; $E = 155$, 187, 219, 251, 280, and 312 V/cm). Figure 5 shows that S increases as electric field strength decreases, demonstrating that sequencing read length is the longest at the highest possible electric field. (The 27mer drag-tag separations follow the same trend; data not shown). The trend of S in Figure 5 follows closely with the actual sequencing read length for these separations; the read length is the longest at 312 V/cm for both the 27mer (210 bases) and 36mer drag-tags (265 bases); read length then decreases with electric field. The increase in band broadening at $E = 312$ V/cm from likely analyte−wall interactions will eventually cause a decrease in sequencing read length when even longer positively charged protein drag-tags are used; however, the slight increase in H at highest E seen in Figures 4A and 4B with net charge of +1 and +2 on the drag-tags is not enough to affect read length, as shown by Figure 5. This data proves that adding net charge up to +2 (total of 4 positively charged amino acids) on protein drag-tags increases friction without added length and still achieves the longest possible sequencing at the fastest possible speed.

α Value. As the effective friction α increases, sequencing resolution of longer fragments is attainable. For this family of protein polymer drag-tags, the α value increased more than 50% with doubled drag-tag length from the 18mer to the 36mer. The sequencing read length increased almost 50% as the drag-tag length was doubled. The increase in α from the 36mer to the 72mer is also more than 50%; therefore, theoretically, the 72mer

protein ($\alpha = 152$) might be able to provide sequence up to 400 bases (with a post-PCR conjugation method; a speculative plot of read length vs α is given in Supporting Information). Ionic strength of the separation buffer also influences α; sequencing read length is the longest with 1.0 X TTE buffer (Supporting Information).

CONCLUSIONS

Free-solution conjugate electrophoresis (FSCE) sequencing read lengths were increased approximately 50% to 265 bases with a highly repetitive, genetically engineered, monodisperse 267-aa protein polymer drag-tag. Four drag-tags based on the amino acid "monomer" Gly-Ala-Gly-Ser-Thr-Gly-Ala (27, 36, 54, and 72 repeat units long) were used for FSCE separations, with net charge of +1, +2, +4, and +6. The 27mer and 36mer were used for sequencing, and separations were diffusion-limited at all but the highest electric field used (312 V/cm). Sequencing read length remained longest at 312 V/cm despite increased band broadening likely from interactions between the capillary walls and the positively charged proteins. The added friction from the positive charges may eventually be balanced with a decrease in read length from analyte−wall interactions, but a net charge of +2 remain advantageous.

Miniaturization of FSCE separations onto glass microchips has been successful for genotyping applications; separation time was decreased more than 90% when performed on microchips.[33] A similar decrease is expected for FSCE sequencing on microchips; sequencing of 265 bases is likely to be achieved in ~3 min. The largest protein drag-tags (54 and 72 repeats of the monomer amino acid block) showed significant monodispersity and the 72mer is predicted to have the friction ($\alpha = 152$) necessary to sequence at least 400 bases and through an entire exon for medical diagnostics. Sequencing attempts with the two largest drag-tags indicate that a post-PCR conjugation method is necessary, and it is under development. Future work enabling FSCE-based Sanger sequencing of entire exons on glass microchips could significantly advance the development of automated, ultrafast microchip sequencing instruments, by eliminating the troublesome requirement of pressure-loading chips with viscous sieving networks.

ACKNOWLEDGMENT

This work was supported by National Institutes of Health grants (NHGRI Grants 5 R01 HG002918-04, 5 R01 HG001970-09, and 1 RC2 HG005596-01) as well as a National Science Foundation Graduate Research Fellowship for J.C.A. The authors declare no competing financial interests and thank Corinne Lusher for her help with experiments and Prof. Gary W. Slater for helpful discussions.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review August 19, 2010. Accepted November 23, 2010.

AC102188P