for assembling splicing sites. The ABI SOLiD and the Illumina GAIIX have not only increased the sequencing length to 50 and 75 bases, respectively, but have also developed methods for sequencing from both ends of the cDNA fragments to help in connecting more distant exons.

Other challenges of RNA-seq are how to distinguish the various start and end sites of RNAs. It is becoming evident that there are often multiple overlapping RNAs encoded from the same genome region, and intron-derived RNAs are recycled to produce functional ncRNAs such as microRNAs. Another source of complexity comes from the secondary processing of mRNAs, which produces shorter, likely functional, RNAs. Thus, protein-coding genes are associated with a plethora of short ncRNAs, including short RNAs associated with promoters[13], transcripts arising around termination sites and even exons. A fraction of these RNAs are produced by a novel cleavage and recapping mechanisms, resulting in capped RNAs that start in the middle of coding exons or in untranslated regions. These naturally truncated RNAs are likely to be ncRNAs that overlap larger mRNAs[13]. Another complication arises from the broad nature of many promoters[14], which produce various capped RNAs from multiple transcription start sites. Technologies that identify the cap structure in such mixtures are needed to distinguish the RNA fragments obtained by RNA-seq. At present RNA-seq does not perform well at unambiguously identifying transcription start sites, and RNA-seq protocols need improvement to simultaneously decipher the long, short and capped RNAs so the RNAs' function can be assessed.

Some of the third-generation sequencers such as those from Pacific Biosciences and Oxford Nanopore—which will be able to read thousands of nucleotides[15] of single cDNAs—may ultimately meet these challenges: their long sequences will quantitatively represent complete RNAs, and the use of tags and linkers that mark cap sites and other modifications will allow an all-in-one determination of transcriptome structure, including start and termination sites and the mapping of regulatory elements such as promoters. The accurate sequence of coding sequences will also help directed cloning of open reading frames in experimental verification of alternative splice isoforms[16,17].

Although many challenges are ahead, the direction is becoming clearer, and I am beginning to wonder if the dark age of the transcriptome is giving way to rays of light.

1. Liang, F. *et al. Nat. Genet.* **25**, 239–240 (2000).
2. Carninci, P. *et al. Science* **309**, 1559–1563 (2005).
3. ENCODE Project Consortium *et al. Nature* **447**, 799–816 (2007).
4. Carninci, P. *et al. Genome Res.* **13**, 1273–1289 (2003).
5. Kapranov, P. *et al. Science* **316**, 1484–1488 (2007).
6. Harbers, M. & Carninci, P. *Nat. Methods* **2**, 495–502 (2005).
7. Sultan, M. *et al. Science* **321**, 956–960 (2008).
8. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. *Nat. Genet.* **40**, 1413–1415 (2008).
9. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. *Nat. Methods* **5**, 621–628 (2008).
10. Cloonan, N. *et al. Nat. Methods* **5**, 613–619 (2008).
11. Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009).
12. Faulkner, G.J. *et al. Nat. Genet.* **41**, 563–571 (2009).
13. Fejes-Toth, K. *et al. Nature* **457**, 1028–1032 (2009).
14. Carninci, P. *et al. Nat. Genet.* **38**, 626–635 (2006).
15. Turner, D.J., Keane, T.M., Sudbery, I. & Adams, D.J. *Mamm. Genome* **20**, 327–338 (2009).
16. Djebali, S. *et al. Nat. Methods* **5**, 629–635 (2008).
17. Salehi-Ashtiani, K. *et al. Nat. Methods* **5**, 597–600 (2008).

Engineered fluorescent proteins: innovations and applications

Michael W Davidson & Robert E Campbell

Despite expansion of the fluorescent protein and optical highlighter palette into the orange to far-red range of the visible spectrum, achieving performance equivalent to that of EGFP has continued to elude protein engineers.

Evolving proteins, evolving tools

During the past decade and a half, intrinsically fluorescent proteins have been under intense evolutionary pressure for 'fitness', not in the wild, but rather for utility in live-cell imaging experiments. This unnatural course of evolution has occurred on the benches of protein engineers around the world who have helped to drive progress in the ever-expanding repertoire of fluorescence imaging technologies.

An underlying theme that has guided advancements in fluorescent protein engineering is that, all other factors being equal, redder is better. It is generally accepted that excitation with longer-wavelength light entails less phototoxicity for the cells or tissue being examined and decreased autofluorescence and scattering. These desirable factors mean that red-shifted fluorophores generally provide improved contrast (owing to decreased background fluorescence) and superior performance in whole-organism imaging (owing to higher tissue 'transparency'). Early efforts to engineer red-shifted *Aequeorea victoria* GFP (avGFP) variants led to the development of enhanced GFP (EGFP) and yellow fluorescent proteins with emission maxima at approximately 507 nm and 529 nm, respectively (versus 508 nm for wild type)[1].

For a time, however, it appeared that fluorescent protein engineering had hit a 'yellow' wall in efforts to red-shift fluorescence emission. Fortuitously, this barrier had already been surmounted by natural evolution, as was revealed in October 1999 with a report that the *Discosoma* sp. mushroom anemone harbored a fluorescent protein homolog, commonly known as DsRed, emitting in the orange-red region (583 nm)[2]. Counterbalancing this favorable shift to the red were several undesirable properties, including oligomerization, 'contamination' by a green component and sluggish chromophore development, which dampened some of the initial enthusiasm.

The discovery of DsRed (and other Anthozoa fluorescent proteins of various hues) had a twofold impact on the

Michael W. Davidson is at the National High Magnetic Field Laboratory and Department of Biological Science, Florida State University, Tallahassee, Florida, USA. Robert E. Campbell is at the University of Alberta, Department of Chemistry, Edmonton, Alberta, Canada.
e-mail: robert.e.campbell@ualberta.ca

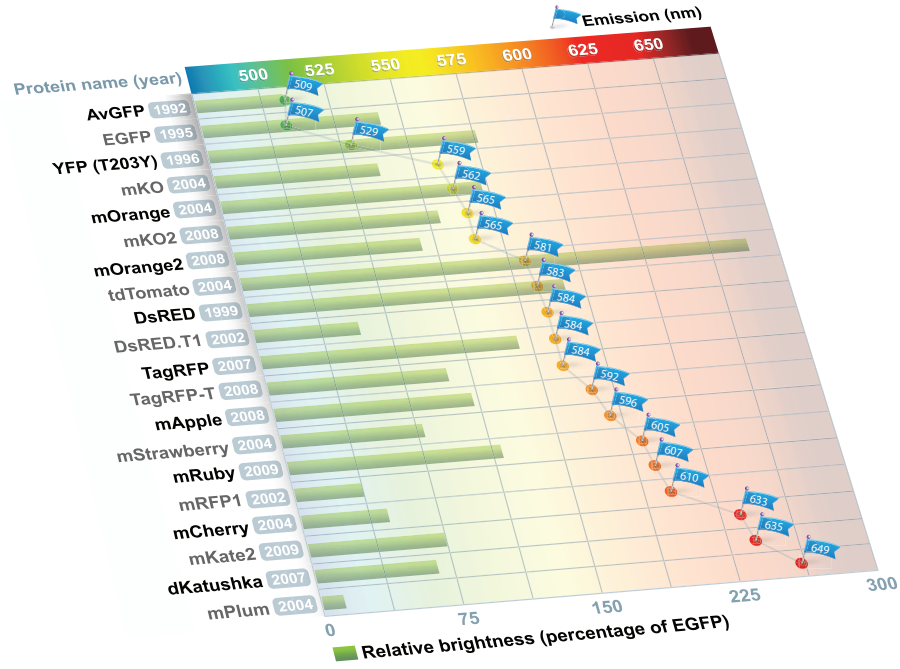

Figure 1 | The fluorescent protein color palette has expanded ~140 nm into longer wavelengths over the past 15 years to include many new variants in the orange, red and far-red spectral regions. These derivatives feature a diversity of properties with respect to brightness, photostability, maturation time and utility in fusions, but none yet match the overall performance of EGFP.

fluorescent protein engineering community. First, there is now a relatively diverse selection of 'templates' from which to undertake directed evolution. Second, recognition of the previously unexpected diversity of spectral properties in the fluorescent protein superfamily goaded protein engineers into trying to push these proteins to new performance extremes in order to provide suitable candidates for imaging at longer wavelengths (**Fig. 1**). These efforts paid off with the development of the so-called 'mFruit' DsRed variants, which are monomeric and fluoresce in hues ranging from orange (mOrange with emission peak at 562 nm) to the far-red (mPlum with an emission peak at 649 nm)[3,4]. Although not as red-shifted as mPlum, a variant known as mCherry is considered to be the preferred choice of the red monomers as it combines red-shifted emission (610 nm), with reasonable photostability, brightness and performance in fusions.

An overriding factor that continues to hamper the choice of an orange or red fluorescent protein for live-cell and *in vivo* imaging is that no single variant performs on par with EGFP in terms of maturation, brightness and photostability. Thus, it may be necessary to use different fluorescent proteins in these spectral regions for particular experiments. Although mCherry remains the optimum monomeric red fluorescent protein for general-purpose imaging, many investigators report problems with aggregation when using mCherry in fusions. Recently, several newer variants have surpassed mCherry in specific areas of performance. For example, mApple outperforms other orange and red fluorescent proteins when fused to connexins targeted at gap junctions, whereas mKusabira Orange2 (mKO2) is a superior fluorescence resonance energy transfer (FRET) acceptor (unpublished observation).

In practice, whether or not a particular fluorescent protein will yield a superior quality of experimental data depends on fundamental (and possibly incidental) details of the particular experimental design. For example, if detection of a sparse target is the goal, a brighter fluorescent protein would be desirable. Accordingly, a good choice might be monomeric TagRFP or the pseudo-monomeric tandem dimer Tomato (tdTomato), which have intrinsic brightness levels that are 2.8- and 6-fold greater than mCherry, respectively, and show excellent performance in many fusions (M.W.D., unpublished observation)[5,6]. On the downside, neither fluorescent protein is as red-shifted nor as photostable as mCherry.

If long-term imaging of fusions is critical, photostability may be of greater importance than other factors. Tsien and co-workers have developed a screen that allowed them to identify red fluorescent protein variants with improved photostability[7]. Their efforts led to the development of the most photostable of the monomeric red fluorescent proteins currently available, including a version of TagRFP known as TagRFP-T and mApple, which has excellent performance in many fusion proteins. These photostable variants benefit from having twice the brightness of mCherry, but they do take somewhat longer to develop red fluorescence and are not quite as red-shifted. Another monomer, named mRuby, is derived from a sea anemone protein and is the brightest red fluorescent protein yet reported with emission above 600 nm, but it has limited photostability and performance in many fusions when compared to mCherry[8].

If the experimental goal is to image as deep into the tissue of a transgenic organism as possible, longer-wavelength excitation and emission is critical. Ideally, both the excitation and emission wavelength would reside in the so-called 'near-infrared window' that extends from 650 to 900 nm. In this region, the combined absorption from hemoglobin and water are minimal and tissue is maximally transparent to light. mPlum was one of the first engineered variants with strong emission at wavelengths greater than 650 nm[4]. Chudakov and co-workers have since reported a dimeric far-red fluorescent protein named Katushka that, owing to an optimal combination of a red-shifted emission peak (635 nm) and high intrinsic brightness, has 7.7-fold greater brightness than mPlum beyond 650 nm[9]. As a dimer, Katushka performs poorly in fusions. The same group reported new derivatives, such as mKate2 and tdKatushka[10], which retain strong emission at wavelengths greater than 650 nm and are very promising for use in chimeras.

Seemingly neglected in these efforts is the fact that even the most red-shifted fluorescent proteins still require excitation well outside of the near-infrared window. An alternate, and perhaps more realistic, solution for imaging deeper into tissue would be the development of a fluorescent protein (or fluorescent protein pair) that has a strong two-photon cross-section at 800–900 nm and strong emission at >650 nm. This could probably be achieved by creating a high-FRET green plus far-red FRET pair.

Far-red fluorescent proteins are particularly promising in the creation of transgenic animals. However, a common problem with many red fluorescent proteins is cytotoxicity. This is an underappreciated and multifaceted problem that protein engineers should spend more time addressing. Such efforts would be greatly facilitated by a systematic investigation of specific causes of fluorescent protein cytotoxicity, the true extent of which is masked by the fact that its reports are mostly anecdotal and are complicated by the vast number of combinations of fluorescent protein variants, fusion partners and cell types involved. In one of the few reported attempts to find a general solution to this problem, Glick and co-workers engineered a DsRed variant with diminished tendency to aggregate and minimal cytotoxicity in a range of cell types[11]. However, to thoroughly address all facets of this problem, researchers may eventually need to resort to an *in situ* process of optimization in which the long-term survival of cells is the selection criterion. One strategy for achieving this might involve multiple rounds of fluorescence-activated cell sorting and passaging of a large library of fluorescent protein variants to select for derivatives that are least detrimental to long-term growth.

With a growing palette of bright fluorescent proteins featuring hues that are continually being pushed to ever–longer wavelengths, a variety of new potential fluorescent protein combinations for use in FRET experiments have become available. Unfortunately, FRET pairs incorporating orange, red or far-red fluorescent proteins have not offered performance on par with the traditional CFP-YFP pair. A common problem when using red-shifted acceptors is the relatively weak sensitized emission because of either a low extinction coefficient or a mismatch in protein maturation rates between the donor and acceptor. Although future developments in FRET partners may overcome these limitations in the longer wavelengths, some investigators have turned their attention back to the blue region of the visible spectrum and developed several new hues that are particularly useful as donors in new FRET pairs[12,13]. The new FRET pairs are sufficiently spectrally distinct to enable simultaneous imaging of dual pairs in a single cell. Keeping the 'redder is better' mantra in mind, we might hope that FRET pairs requiring high-energy violet excitation will one day be made redundant by an orange-red FRET pair with comparable performance.

Photochemistry: boon or bust?

Much as the fluorescent protein palette started with a single green variant and diversified into an entire spectrum of useful colors, so too has the toolbox of fluorescent protein–based optical highlighters continued to expand and improve. The progenitor of this diverse family is photoactivatable GFP (PA-GFP), a variant of avGFP that dramatically increases its fluorescence when illuminated with intense ~400-nm light[14]. Although initially used for selective highlighting of spatially confined protein subpopulations, optical highlighters have seen a resurgence in popularity for use in so-called 'super-resolution' fluorescence imaging.

The development of new optical highlighters has been driven by many of the concerns and considerations (photostability, brightness and others) discussed above. Accordingly, it is apparent that a red analog of PA-GFP would be highly desirable both because of red-shifted emission and as a second color for two-color photoactivation experiments. To address this need, Verkhusha and co-workers converted mCherry into a variant, known as PA-mCherry1, which is initially nonfluorescent and emits red light only after illumination with violet light[15]. PA-mCherry1 enabled a two-color implementation of the super-resolution imaging technique known as photoactivated localization microscopy, or PALM. The combination of PA–mCherry1 and PA-GFP also has excellent potential for use in two-color single-molecule tracking in live cells. The Achilles' heel of PA-GFP and PA–mCherry1 is limited brightness and difficulty in locating regions of interest for activation. Clearly, there is substantial need for improvements in this fluorescent protein class.

As with other fluorescent proteins, optical highlighters have been subjected to extensive engineering modifications that include monomerization as well as optimization of brightness and performance in fusion chimeras. One of the recent developments in this arena is the bright and monomeric (weakly dimeric) mEos2 protein[16]. mEos2 is an example of a photoconvertible fluorescent protein that can be irreversibly switched from green to red emission upon illumination with violet light. Owing to a combination of high brightness and photostability, mEos2 enables PALM imaging with high localization precision.

It is clear that fluorescent proteins have become an established and trusted tool. However, a critical lesson to be gleaned from the development of fluorescent protein–based highlighters is that many fluorescent proteins can undergo fairly complex photochemistry upon excitation. It has been observed that, under certain intense illumination conditions, YFP can be photoconverted into a hue-shifted species very similar in spectral properties to CFP[17]. This phenomenon can be troublesome because live-cell FRET efficiency is often determined by photobleaching of the acceptor (often YFP) and quantifying the increase in donor (often CFP) signal. The YFP to CFP photoconversion artifact could potentially cause an investigator to overestimate the FRET efficiency or to observe apparent FRET when none was actually present. Photoswitching and photoconversion also seem to be general properties of red and orange fluorescent proteins[18], prompting the need to be wary of spontaneous fluorescence recovery or shifts to new emission wavelengths during photobleaching experiments. Fortunately, these issues have now been identified, and awareness of potential artifact sources is becoming general knowledge in the research community.

Fluorescent proteins of the future

The currently available fluorescent proteins emitting in the orange, red and far-red wavelengths do not feature similar performance with regard to utility in fusions, brightness and photostability as EGFP. Furthermore, some of the most aggressively optimized far-red fluorescent proteins retain a substantial green-fluorescent component, essentially making them impractical for two-color imaging with EGFP. Clearly, a modest and realistic near-term goal for the community should be the development of a red fluorescent protein that matches EGFP in all performance aspects. In addition, efforts to push red fluorescent protein variants to new extremes of red-shift, brightness and photostability should continue, but with the expectation that all of these favorable properties may be mutually incompatible.

We also face the conundrum of how to engineer a fluorescent protein with near-infrared absorption and emission if the inherent potential for red-shifting the fluorescent protein chromophores has been exhausted. One possibility is that nature has already solved this problem and there exists an unidentified reef organism with a fluorescent protein homolog emitting in the near-infrared. In the absence of such a fortuitous discovery, protein engineers must search for

alternative chromophores with extended conjugation in order to access the near-infrared. Perhaps the solution is to abandon the fluorescent protein superfamily entirely and look to alternative protein scaffolds for development of near-infrared fluorescent proteins. Such an approach was recently taken by Tsien and co-workers in the development of an infrared fluorescent protein based on a bacteriophytochrome protein that binds an ubiquitous biliverdin chromophore[19].

The rise of fluorescent proteins as powerful imaging tools has been marked by fairly regular surprising revelations regarding the versatility and utility of these remarkable fluorescent probes. What other surprises do fluorescent proteins have in store? Although we cannot answer that question with any degree of certainty, one thing is clear: when fluorescent proteins next reveal an interesting new property, the research community will once again seize it and exploit it to enable exciting new applications on the cutting edge of biological imaging technology.

COMPETING INTEREST STATEMENT
The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

1. Tsien, R.Y. *Annu. Rev. Biochem.* **67**, 509–544 (1998).
2. Matz, M.V. *et al. Nat. Biotechnol.* **17**, 969–973 (1999).
3. Shaner, N.C., Steinbach, P.A. & Tsien, R.Y. *Nat. Methods* **2**, 905–909 (2005).
4. Wang, L., Jackson, W.C., Steinbach, P.A. & Tsien, R.Y. *Proc. Natl. Acad. Sci. USA* **101**, 16745–16749 (2004).
5. Merzlyak, E.M. *et al. Nat. Methods* **4**, 555–557 (2007).
6. Shaner, N.C. *Nat. Biotechnol.* **22**, 1567–1572 (2004).
7. Shaner, N.C. *et al. Nat. Methods* **5**, 545–551 (2008).
8. Kredel, S. *et al. PLoS One* **4**, e4391 (2009).
9. Shcherbo, D. *et al. Nat. Methods* **4**, 741–746 (2007).
10. Shcherbo, D. *et al. Biochem. J.* **418**, 567–574 (2009).
11. Strack, R.L. *et al. Nat. Methods* **5**, 955–957 (2008).
12. Ai, H., Hazelwood, K.L., Davidson, M.W. & Campbell, R.E. *Nat. Methods* **5**, 401–403 (2008).
13. Tomosugi, W. *et al. Nat. Methods* **6**, 351–353 (2009).
14. Patterson, G.H. & Lippincott-Schwartz, J. *Science* **297**, 1873–1877 (2002).
15. Subach, F.V. *et al. Nat. Methods* **6**, 153–159 (2009).
16. McKinney, S.A., Murphy, C.S., Hazelwood, K.L., Davidson, M.W. & Looger, L.L. *Nat. Methods* **6**, 131–133 (2009).
17. Valentin, G. *et al. Nat. Methods* **2**, 801 (2005).
18. Kremers, G.J., Hazelwood, K.L., Murphy, C.S., Davidson, M.W. & Piston, D.W. *Nat. Methods* **6**, 355–358 (2009).
19. Shu, X. *et al. Science* **324**, 804–807 (2009).

Comparative analysis to guide quality improvements in proteomics

Matthias Mann

The potential of mass spectrometry–based proteomics to advance biology and biomedicine is nearly unlimited but so is its potential for generating bad data. Apart from the pursuit of technological progress in protocols and instruments, stringent comparative analyses of different approaches are critical for fully developing the discipline.

Systems-wide analysis of any biological entity is hard. For proteins it is particularly daunting because proteomics lacks an equivalent to hybridization assays based on Watson-Crick base pairing or amplification reactions based on PCR. Fortunately, we do have the wonderful tool of mass spectrometry (MS) at our disposal. MS is one of the most versatile technologies used in biology, and continuing radical improvements in the technology amaze even the most seasoned observers. However, not all is well in the discipline of proteomics, and much fuzzy thinking and bad data have unfortunately found their way into the literature. Purely technological improvements in proteomics will go a long way to overcome these difficulties[1], but they need to be accompanied by rigorous analysis. Three 'analysis' papers published in the last five years in *Nature Methods* have broken new ground in investigating crucial data quality issues in the proteomics field. One of them deals with protein identification, another with the enrichment of phosphorylated peptides, and the third with the evaluation of proteomics researchers themselves!

A main developmental direction of our discipline is pushing the identification of ever more proteins in complex proteomes, something to which MS is uniquely suited. However, many of the early landmark papers in the last 5–10 years that established the feasibility of large-scale protein identification were obtained on low-resolution instruments and without proper statistical analysis. We now know that a large proportion of the identifications obtained from such projects were in fact false positives. For example, peptide lists contained a large proportion of nontryptic peptides, whereas it is now generally acknowledged that trypsin, at least in proteomics experiments, is extraordinarily sequence-specific[2]. The recognition of these data quality issues prompted a gradual, though still not complete, switch to high-resolution techniques. It also lent impetus to efforts to standardize the reporting of proteomics protocols and data, and to the development of bioinformatics tools to directly determine the false positive rate independently of the peptide database search score.

Aebersold and co-workers addressed this issue early on by developing an algorithm that decomposed scored distributions into underlying false positive and true positive distributions and assigned likelihoods that peptides with a given database identification score were in fact correctly identified. This development, implemented in their PeptideProphet and ProteinProphet software[3,4], was an important step in bringing some rigor to the identification process in low-resolution data. Even more simple and powerful—and applicable to high- as well as low-resolution data—was the concept of applying reverse sequence databases to determine false positive identification rates. This approach is very straightforward and only involves searching the data against the normal or 'forward' database and against the sequence-reversed database (also called a 'target-decoy' database, if it is

Matthias Mann is at the Max Planck Institute for Biochemistry, Matrinsried, Germany.
e-mail: mmann@biochem.mpg.de