Bioinformatics

Finding genes in *Plasmodium falciparum*

The completion of the sequencing of chromosome 3 of the malarial parasite *Plasmodium falciparum*[1] is a major step forward in our understanding of the *Plasmodium* genome. We have analysed this chromosome using GlimmerM, a freely available gene-finder developed specifically for the *P. falciparum* species[2]. GlimmerM was highly effective in finding nearly all the genes that were reported on *P. falciparum* chromosome 2 (ref. 2), and a newly retrained version was even more effective on chromosome 3, confirming virtually all reported genes and finding several additional ones.

Using GlimmerM, we found 25 possible new genes on chromosome 3 in regions currently annotated as non-coding[1] (Table 1). Although some of these are relatively short and may not be genuine coding sequences, six of them are longer than 400 base pairs (bp) and one is 2,187 bp. This last gene represents a 729-amino-acid protein that is encoded by one very large exon and one shorter exon. Open reading frames of this length in a chromosome with 80% A+T content are virtually certain to represent real genes. Three of the table entries, G802, G803 and G740, have detectable homology to *var* gene fragments on chromosome 2; G802 and G803 are close enough that they might represent two portions of the same gene.

Of the 215 protein-coding regions reported by Bowman *et al.*[1], GlimmerM automatically finds 214 of them. (An earlier version of GlimmerM was provided to the annotators of chromosome 3.) The only gene missed is a short hypothetical protein (PFC0360w, 114 amino acids) with no homology to any known gene[1]. Finally, given that chromosome 3 is 12 per cent larger than chromosome 2, an extrapolation based on gene density would predict 234 genes on chromosome 3, consistent with our finding that additional genes could be present.

GlimmerM is very accurate at identifying splice sites, having been trained on a carefully curated set of experimentally confirmed introns from the *P. falciparum* genome[2]. GlimmerM's predictions suggest different splice sites for 49 of the 215 genes annotated on chromosome 3; all of these are hypothetical proteins. As with the annotation of hypothetical proteins for chromosome 2, substantial additional laboratory studies are needed in order to determine with confidence the exon structure of these genes.

For chromosome 2, we used the polymerase chain reaction with reverse transcription for 13 hypothetical genes predicted by GlimmerM: all 13 predictions were confirmed[2,3]. Of course, as emphasized previously[2,3], GlimmerM is only one step, albeit an important one, in a process that should involve many other computational methods as well as careful human curation of the results produced by those methods.

To achieve the highest quality in analysing genome sequences, peer-reviewed bioinformatics methods should be used when they are available. We consider it an oversight that the chromosome 3 annotation effort neglected to consider the predictions of GlimmerM, especially as a substantial part of the chromosome 3 analysis involves comparing the two chromosomes (see, for example, Table 2 of Bowman *et al.*[1]).

Mihaela Pertea, Steven L. Salzberg, Malcolm J. Gardner

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA
e-mail: salzberg@tigr.org

1. Bowman, S. *et al. Nature* **400,** 532–538 (1999).
2. Salzberg, S. L. *et al. Genomics* **59,** 24–31 (1999).
3. Gardner, M. J. *et al. Science* **282,** 1126–1132 (1998).

Table 1 Possible new genes on chromosome 3

ORF	No. of exons	Length		Flanking genes		Strand
		BP	AA			
G802	1	450	150	PFC0010c	PFC0025c	–
G803	1	468	156	PFC0010c	PFC0025c	–
G815	1	210	70	PFC0030c	PFC0035w	+
G036	2	192	64	PFC0125w	PFC0130c	–
G070	1	220	73	PFC0175w	PFC0180c	–
G121	1	531	177	PFC0260w	PFC0265c	–
G124	1	411	137	PFC0260w	PFC0265c	–
G141	1	255	85	PFC0280c	PFC0285c	+
G144	4	348	116	PFC0280c	PFC0285c	+
G233	1	240	80	PFC0415c	PFC0420w	+
G256	1	459	153	PFC0440c	PFC0445w	–
G273	1	282	94	PFC0465c	PFC0470w	–
G282	2	249	83	PFC0480c	PFC0485w	–
G290	1	243	81	PFC0485w	PFC0490w	–
G294	1	285	95	PFC0490w	PFC0495w	+
G313	2	285	95	PFC0505c	PFC0510w	+
G356	2	293	131	PFC0555c	PFC0560c	–
G408	2	213	71	PFC0580c	PFC0590c	+
G410	1	231	77	PFC0580c	PFC0590c	+
G529	5	339	113	PFC0810c	PFC0815c	–
G612	3	249	83	PFC0910w	PFC0915w	–
G614	2	279	93	PFC0910w	PFC0915w	+
G702	2	2187	729	PFC1010w	PFC1015c	–
G734	1	237	79	PFC1060c	PFC1065w	+
G740	1	312	104	PFC1065w	PFC1070c	–

Analysis of chromosome 3 of *P. falciparum* by GlimmerM[1] generated 25 new gene models not included in current annotation. Columns show the number of exons in the predicted gene model, the length in base pairs (BP) and amino acids (AA), the nearest flanking genes on either side of the predicted new gene and the strand (+, forward strand; –, reverse complement). A version of this table, linked to the DNA and protein sequences for each gene, is available at http://www.tigr.org/tdb/edb/pfdb/pf3table.html. ORF, open reading frame.

Lawson et al. *reply* — Pertea *et al.* report their interpretation of the published *Plasmodium falciparum* chromosome 3 sequence[1] using The Institute for Genomic Research's gene-prediction algorithm GlimmerM[2]. We believe, however, that their analysis is flawed, and highlights the problem that overreliance on a single algorithm can lead to mistakes in annotation.

The accurate *ab initio* prediction of a eukaryotic coding sequence is extremely difficult because of the presence of intronic DNA. Gene predictions are distillates of the results from several algorithms that differentiate coding from non-coding segments. The role of the annotator is to identify putative genes and achieve the best prediction using the tools available. Ultimately, predictions can only be validated by laboratory studies.

GlimmerM finds 25 new gene models, which Pertea *et al.* suggest could be an oversight in the original analysis. Predictions G802, G803 and G740 map to regions of chromosome 3 already assigned as varC pseudogenes, as the region of similarity is heavily frameshifted[1]. We do not believe that the prediction of exons from within these regions correctly interprets the data. Until laboratory investigation can show that these regions are coding, the parsimonious (and correct) annotation is that these regions are pseudogenes.

Many gene-prediction algorithms are designed to have a minimum gene size (quite often 100 amino acids). These limits are implemented not because of any biological rationale, but because the inclusion of such open reading frames presents major difficulties in the interpretation of the data as most are likely to be non-coding. There is

no indication that the 15 novel small predictions are coding, and the hexamer algorithm suggests that 12 of the 15 are non-coding. The remaining seven GlimmerM predictions form part of gene models that we consider to be borderline and are validating experimentally before inclusion in the chromosome annotation.

Pertea *et al.* suggest that peer-reviewed bioinformatics methods are essential. Certainly, if the diverse outputs of multiple gene-prediction algorithms can be included, then the annotation should be good, but that does not make the gene prediction correct — it remains a prediction and not a confirmed gene model.

Overreliance on the output of a predictive tool can lead to erroneous prediction and bad annotation. The well-established tools used in the analysis of chromosome 3, Genefinder, hexamer and ACEDB (P. Green and L. Hillier, unpublished software; R. Durbin, unpublished software) have a proven track record in eukaryotic gene prediction[3].

Our analysis of chromosome 2 using Genefinder/hexamer indicates that 40 modifications need to be made to the original gene predictions (20 per cent of gene models). Although these new predictions require confirmation, they highlight the inaccuracy of first-pass annotation from uncharacterized genomic DNA. We encourage re-analysis of genomic data to increase sequence accuracy: Genefinder, hexamer and GlimmerM all have roles to play in the (re)annotation of *P. falciparum* chromosomes and in improving the interpretation of sequence data.

Dan Lawson, Sharen Bowman, Bart Barrell
Pathogen Sequencing Unit, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK
e-mail: dl1@sanger.ac.uk

1. Bowman, S. *et al. Nature* **400,** 532–538 (1999).
2. Salzberg, S. L. *et al. Genomics* **59,** 24–31 (1999).
3. The *C. elegans* Sequencing Consortium *Science* **282,** 2012–2018 (1998).

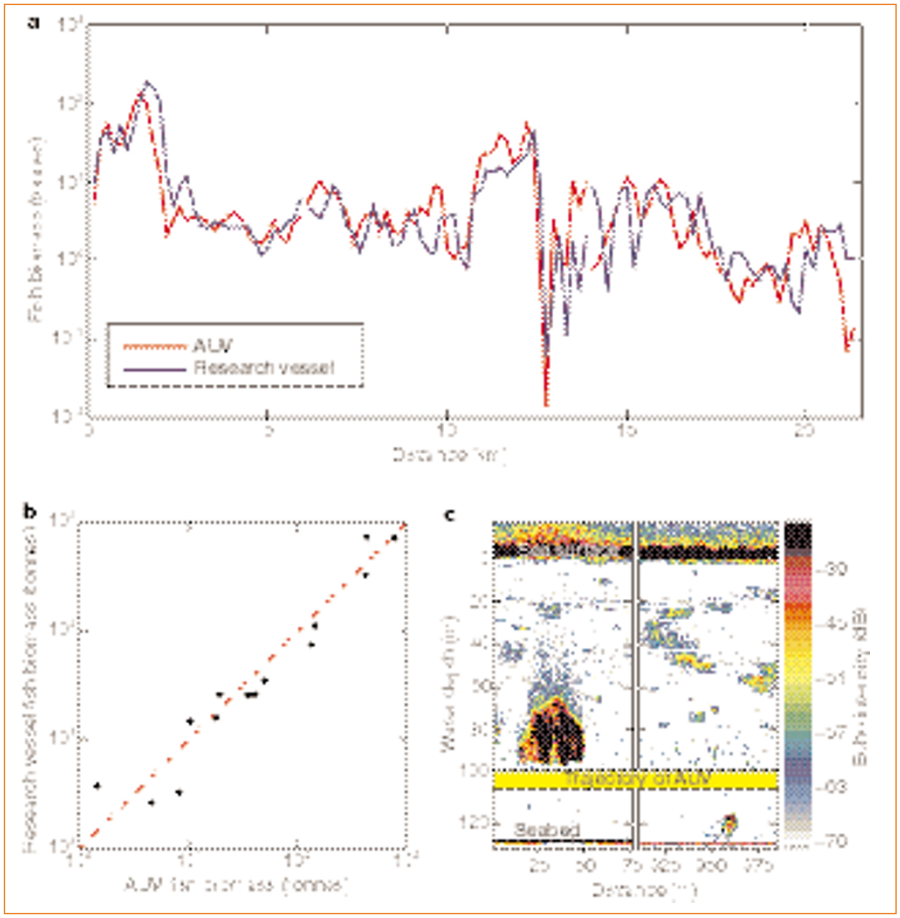

Figure 1 Comparison of acoustic data collected by an autonomous underwater vehicle (AUV) and a research vessel. **a,** Fish biomass at integrated intervals of 186 m along three transects on 21 July 1999; fish were identified as herring (*Clupea harengus*) by trawling. **b,** Strata-averaged biomass estimates from the AUV and research vessel were significantly correlated ($r = 0.935$, $P < 0.001$), lying around the (dotted) one-to-one line (Wilcoxon signed-rank test, $P > 0.05$). **c,** Composite echogram collected by the AUV at 105 m (equipped with upward- and downward-looking transducers) as it passed less than 7 m beneath a large midwater herring school (left) and only 10 m above another close to the sea bed (right).

Oceanography

Fish do not avoid survey vessels

The precarious condition of the world's fisheries is making ever-greater demands of the scientific assessment of fish stocks. Traditional assessments that rely on commercial catch statistics can have major shortcomings[1] (as shown, for example, by the collapse of Canada's northern cod stock[2]), increasing the need for more fishery-independent data. Acoustic surveys can provide such information[3], but ocean-going research vessels have high operating costs, and there is also widespread concern that fish avoid these vessels because of the noise they make, thereby biasing abundance estimates[4]. Here we present new data gathered by an autonomous underwater vehicle (AUV) showing that vessel avoidance is not a significant source of bias. Our investigation also heralds the arrival of AUVs as effective survey platforms.

During acoustic surveys, sound pulses are transmitted vertically downwards into the water at regular intervals (typically 1 s) from a survey vessel travelling along defined transects. Fish density is calculated by integrating the intensities of the returning echo[5], and is then interpolated to give an estimate of the abundance in the survey area[3].

We deployed the AUV *Autosub-1* (ref. 6) 200–800 m ahead of the research vessel *Scotia* on eight transects in water 60–180 m deep during an acoustic survey of herring in the North Sea. Herring have the most sensitive hearing of the commercially exploited species examined so far[7] and are more likely than any to react to vessel noise. *Autosub-1* is unmanned and follows pre-programmed mission trajectories. In comparison to the 68-m *Scotia*, it is small (torpedo shaped, 7 × 1 m) and extremely quiet (being propelled by an electric motor).

Avoidance of *Autosub-1* by herring is minimal: passing unprecedently close to a school (Fig. 1c), the vehicle caused only the localized school compression that typically occurs on close approach of predators[8]. *Autosub-1* was equipped with the same type of 38-kHz scientific echosounder as *Scotia*, and gathered equivalent acoustic data before the research vessel arrived. If fish avoided the *Scotia*, then it should have detected fewer fish than *Autosub-1*.

At the integrated resolution, there are small-scale temporal and spatial differences between the AUV and research-vessel data (Fig. 1a), but the underlying similarity in magnitude and trend is clear. For statistical comparison, data from each source were aggregated into independent strata of equivalent geographical area.

The amount of fish detected by the research vessel was not significantly different from that detected by the AUV (Fig. 1b). *Scotia* is very quiet, having been built