

## Microarray Analysis Characterizes Function of 95% of Malaria-causing *Plasmodium falciparum* Genes, Three Times More than Ever Before

The Scripps Institute's Elizabeth Winzeler and the Sanger Institute's Céline Carret discuss recent advances in malaria research made possible using whole-genome *Plasmodium* microarrays

By Stacey Ryder

**PHILADELPHIA, February 1, 2006** — Scientists at the Scripps Institute and the Novartis Research Foundation have started to uncover the function of more than 95 percent of the malaria-causing *Plasmodium falciparum* genes—three times more than ever before. They're now focusing on 100 of those genes involved in the parasite's

life cycle as targets for new antimalarial drugs and vaccines.

In fewer than 12 months, the team measured the expression of 5,159 genes during nine stages of the parasite's life cycle using a *Plasmodium* microarray. Traditional gene-by-gene methods of functional characterization

in *Plasmodium* have been labor-intensive and time-consuming, and efforts to clone large numbers of genes have been hindered by the AT-rich genome.

"I think our microarray data is having a very positive impact on the way people do research," said Winzeler, associate professor of cell biology at

**Elizabeth Winzeler** is an associate professor

in the Department of Cell Biology at the Scripps Research Institute and leader of the microbiology group at the Genomics Institute of the Novartis Research Foundation. Her group is interested in using large-scale data sets and high-throughput approaches to catalog the malaria genome. Their goal is to be able to predict function for many uncharacterized genes in the *Plasmodium falciparum* genome and to discover genes that are good targets for drug therapies.



the Scripps Institute. “Instead of spending time doing a northern, people can just look at the database. They’ve saved the time that they would have spent on this step and now they can do transfections, westerns or protein expression immediately.”

Céline Carret, a postdoc at the Wellcome Trust Sanger Institute, is investigating deletions and polymor-

so we used some of the gene models that the PlasmoDB supplied.

PlasmoDB ran some of the gene prediction algorithms on the raw sequence data, so we had ideas about which regions were potential coding regions and which were non-coding. Since we were a bit limited in the number of probes we could choose in 2001 or 2002, we biased our selection

### “Mismatchless” microarrays

**Carret:** Given that you have just perfect match probes on these malaria arrays, enabling you to place twice as many features on the array from both strands of the genome, do you think it could become a widespread application for other users?

**Winzeler:** For me, the choice was very clear. I would much rather have those extra functional probes on the chip than the mismatches. Talking to other people, it sounds like mismatches may produce slightly greater sensitivity at the low end of the hybridization spectrum, but I haven’t seen any evidence in my work that they would improve things significantly.

There’s no evidence that using mismatches would improve, for example, what we consider background. It’s just so much more useful for me to have additional redundancy in terms of the number of probes on the chip than having mismatches.

I would certainly encourage people to think about doing away with the mismatches. The only problem with not having mismatches on the array is that people can’t really use it with conventional Affymetrix software—only laboratories that have a fairly good bioinformatics support system have been able to use it.

I haven’t really encouraged the use of our chips in laboratories that just have a microarray facility that uses canned software. But maybe we’ll get past this problem in the future and Affymetrix will develop algorithms that allow the analysis of mismatchless chip data. But by that time, I think this chip will probably be obsolete.

**Carret:** So, you don’t think the lack of mismatches affects the statistical analysis for the perfect matches?

**Winzeler:** If it does, the effect is very small. I don’t think it makes much of a difference. Maybe it’s more of an issue in humans, where you’re hybridizing a much more complex genome. In fact,

“So, even if it isn’t giving us a drug tomorrow, I hope it will at least accelerate the development of new drugs.”

phisms associated with Plasmodium disease pathology using a microarray that tiles probes, on average, every five base pairs (the new PFSanger array). They used these microarrays along with novel amplification methods to examine small amounts of genomic DNA derived from non-laboratory clones of Plasmodium.

Winzeler spoke to Carret about the challenges and successes they have both had in creating microarrays for malaria research. They discussed:

- probe selection and design of Plasmodium microarray
- choosing between the MOID and RMA algorithms for data analysis
- advances that might accelerate malaria research and the potential impact on malaria cases worldwide

### Design of *Plasmodium falciparum* and *scrMalaria* genome microarray

**Carret:** Did you come up with the design for the malaria chip yourself or did you do it in collaboration with Affymetrix?

**Winzeler:** We did everything ourselves. We just sent Affymetrix the sequences of the probes that we wanted to place on the chip. We did this work before the annotated genome was available,

toward coding regions. Our chip has significantly more probes in the coding regions.

The other reason that we don’t have a great number of probes in the non-coding regions is that some of these regions are extremely AT-rich, and it’s very difficult to select functional probes in these regions using probe selection rules simply because they might be long stretches of A’s and T’s. I would like to have more probes in the intergenic regions, but I don’t know how many more functional probes you could get or whether the low melting temperatures they’d have would affect the performance of those probes. It’s something I’m eager to find out.

At the Sanger Institute, you have created a new chip that is essentially a tiling array, right?

**Carret:** Well, it’s not a 100 percent tiling array for exactly the same reason you’ve been talking about—the very AT-rich regions. Near the centrosome, it’s so AT-rich that there’s no way we could design anything we could use. But we have pretty good coverage in the intronic and intergenic regions. Maybe it’s just because it’s four years later and the tools are better now. I don’t know.

some of the statistical analyses that we've run suggest that you're actually adding noise by doing the subtraction because you've added another mathematical step into it, and that reduces your confidence.

**Carret:** That's a good point.

### Comparing the MOID and RMA algorithms

**Carret:** That brings me to MOID, the match-only algorithm you use for the analysis in the *Science* paper. I recently saw a paper by Ken Simpson and Terry Speed where you were a co-author. In that paper, they said that MOID is not as good as RMA for analysis of only perfect matches. What do you think about that?

**Winzeler:** I think the RMA is a really interesting algorithm and it may be that it performs better under some circumstances. Ken did show that there were some incremental improvements over MOID. On the other hand, I really don't care that much about whether my gene is changing by 1.5- or 1.8-fold. It's just not where I want to put my time and energy.

Ultimately, the best test of the two algorithms would be to examine how well each one finds correlation between co-expressed genes. We know that there are a number of genes that seem to be very tightly correlated in their expression patterns, for example, ribosomal genes. You could run the two different algorithms and determine whether you find tighter clustering with RMA or MOID. That would be the most convincing test for me.

I'm still using the MOID. That's partly because we continually add data to our sets. To change the algorithm we'd have to go back and recode everything and reanalyze all of our old data and so forth. It's just a huge amount of work. MOID seems to work pretty well for us. I've been very pleased with the results.

**Carret:** Terry Speed and Ken Simpson agreed with you about MOID. In their paper, they said that the more expressed the gene is, the better it is represented by MOID compared to RMA. RMA tends to decrease the vari-

ance. But RMA is really good at detecting low expression.

So, probably for life cycle work, it's good to see big variation in the signal intensity, but if, for instance, you want to see the effect of some drug on *P. falciparum* cultures or the effect of some knockout genes on the genome expression profile, where you're not sure that it will be a massive change, and you are expecting indeed a small difference in few genes, maybe RMA is more powerful.

**Winzeler:** That may be perfectly true, and certainly we've done expression analysis on some conditions where we've seen very few changes, and maybe those would be really good examples where we should go in and try the RMA algorithm and see if our data looks better under those circumstances. But I think it would still be useful to continue to run MOID on the data just to get estimates of transcript abundance.

**Céline Carret** is a postdoctoral fellow at the Wellcome Trust Sanger Institute. Her work is focused on developing whole-genome tiled microarrays for *Plasmodium falciparum*. She is particularly interested in using microarrays to detect the genomic variation in *P. falciparum* isolates that affects virulence and disease.



## Protocols and validation for an AT-rich genome

**Carret:** If you wanted to look at more intergenic regions, which are, of course, more AT-rich, would you need to make changes to the protocol recommended by Affymetrix?

**Winzeler:** You might get more data by lowering the hybridization temperature, because some of these very AT-rich regions, the target may not stick as well, and you may have poor hybridization intensity. But I haven't tested that.

In general, the AT-rich regions may make it difficult to clone *Plasmodium* genes, but from what I hear from people who have worked with Affymetrix chips, an AT-rich genome generally gives better hybridization data. You have less secondary structure than you have with GC-rich genomes. And the poorest quality data comes from those genomes with huge numbers of G's and C's.

**Carret:** Did you validate what you found on the microarrays using real-time PCR or something like that?

**Winzeler:** Yes. We've done some of that. Having watched people work with RT-PCR, my feeling is that the error margin is significantly higher than with the chip, so I'm not really sure it's a good validation tool.

I've seen a postdoc who has very good hands and can get 99.9 percent correlation with chip hybridizations go and try to do RT-PCR. She'll get different results from one day to the next. It's very frustrating to have that as a gold standard when I believe we have extremely high-quality microarray data.

I personally think statistical methods, such as finding co-expressed genes, finding transcriptional regulatory networks or motifs associated with co-expressed genes or validation in two hybrid sets, provide better validation than randomly selecting one or two genes and then doing RT-PCR.

## Future research possibilities

**Carret:** Given that an increased efficiency transfection system has now been developed for *Plasmodium*, are you contemplating a systematic gene-by-gene knockout project that would be analyzed on the chips?

**Winzeler:** I think that would be a great project. The bottleneck is not in actually doing the transfections, but in



Elizabeth Winzeler

creating the constructs for doing the transfections. That's where you can really spend your time.

If we could reduce the cloning bottleneck by developing ways to make synthetic genes, that would be really exciting. And if we could put tags and so forth into some of these strains, and follow them in mixed populations, I think that would be interesting as well.

We could also try to develop a transposon mutagenesis strategy. I think John Adams mentioned that he might have some transposons working for *Plasmodium*.

So, I think those are really interesting things that we should think about. The funding agencies should think about them, too. And I'd like to be involved if I could. We're just starting to do some transfections in our laboratory now.

**Carret:** Do you think that the data that we can obtain from microarrays could help the children dying of malaria in Africa?

**Winzeler:** I think our microarray data is having a very positive impact on the way people do research. Instead of spending time doing a northern, people can just look at the database. They've saved the time that they would have spent on this step and now they can do transfections, westerns or protein expression immediately.

The data's only been out for a year and a half and it's been cited 155 times already. I think it is currently the most highly cited malaria paper from 2003. So people are paying attention to it, and I've really been delighted when I hear people say, "Well, that data's been very good. It agrees exactly with what we observed." So, even if it isn't giving us a drug tomorrow, I hope it will at least accelerate the development of new drugs.

What I'd really like to do is just to give our chips, or maybe have the Sanger Institute find a way to distribute the newer ones that are coming out, to repositories, like MR4. And just give them a piece of code that they could use that would allow us to transfer the technology to other malaria laboratories that don't have quite the bioinformatics expertise that we and/or the Sanger Institute have.

## AFFYMETRIX MICROARRAY BULLETIN

### Editorial Staff

Wes Conard, *Editor-in-Chief*  
wes\_conard@affymetrix.com  
Tommy Broudy, *Managing Editor*  
thomas\_broudy@affymetrix.com  
Rachel Shreter, *Editor*  
rachel\_shreter@affymetrix.com  
Kamalia Dam, *Associate Editor*  
Stacey Ryder, *Associate Editor*  
Daniel Noble, *Copy Editor*  
Jim Butler, *Contributing Designer*

**Contacts**

■ Elizabeth Winzeler, Ph.D.  
Associate Professor  
Department of Cell Biology  
The Scripps Research Institute  
10550 N. Torrey Pines Rd.  
ICND 202  
La Jolla, CA 92037  
winzeler@scripps.edu

■ Céline Carret, Ph.D.  
The Wellcome Trust Sanger Institute  
Wellcome Trust Genome Campus  
Hinxton  
Cambridge  
CB10 1SA  
United Kingdom  
ckc@sanger.ac.uk

**Companies**

■ Affymetrix Inc. – <http://www.affymetrix.com>

**Organizations**

■ The Scripps Research Institute –  
<http://www.scripps.edu>

■ Genomics Institute of the Novartis Research  
Foundation – <http://web.gnf.org>

■ The Wellcome Trust Sanger Centre –  
<http://www.sanger.ac.uk>

**Further Reading**

■ Carret C, Horrocks P, Konfortov B, Winzeler EA, Qureshi M, Newbold C, Ivens A. Microarray-based comparative genomic analyses of the human malaria parasite *Plasmodium falciparum* using Affymetrix arrays. *Mol Biochem Parasitol.* 2005; 144(2):177-186.

■ Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De la Vega P, Holder AA, Batalov S, Carucci DJ, Winzeler EA. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science.* 2003;301(5639):1503-1508.

■ Young J, Fivelma QL, Blair PL, De la Vega P, Le Roch KG, Zhou Y, Carucci DJ, Baker DA, Winzeler EA. The *Plasmodium falciparum* sexual development transcriptome: A microarray analysis using ontology-based pattern identification. *Mol Biochem Parasitol.* 2005;143(1):67-79.

■ Kidgell C, Winzeler EA. Elucidating genetic diversity with oligonucleotide arrays. *Chromosome Research.* 2005;13:225-235.

■ Simpson KM, Baum J, Good RT, Winzeler EA, Cowman AF, Speed TP. A comparison of match-only algorithms for the analysis of *Plasmodium falciparum* oligonucleotide arrays.

*Int J Parasitol.* 2005;35(5):523-531.

**People**

■ Ken Simpson, Ph.D.  
The Walter and Eliza Hall Institute  
Department of Genetics and Bioinformatics  
<http://www.wehi.edu.au/research/overview/gbi.htm>

■ Terry Speed, Ph.D.  
The Walter and Eliza Hall Institute  
Department of Genetics and Bioinformatics  
<http://www.wehi.edu.au/research/overview/gbi.htm>

■ John Adams, Ph.D.  
The Center for Tropical Disease Research and Training  
[http://ctdrt.bio.nd.edu/index.php?content=member\\_info.php&cid=15](http://ctdrt.bio.nd.edu/index.php?content=member_info.php&cid=15)

**Databases**

■ PlasmoDB – <http://www.plasmodb.org>

■ Malaria Research and Reference Reagent Resource Center (MR4) – <http://www.malaria.mr4.org>