

Genome-wide SNP detection of *Plasmodium vivax* Peruvian patient isolates reveals highly polymorphic genes

Scott Westenberger – TSRI, Winzeler Lab

Scott Westenberger – TSRI, Winzeler Lab

highly polymorphic genes

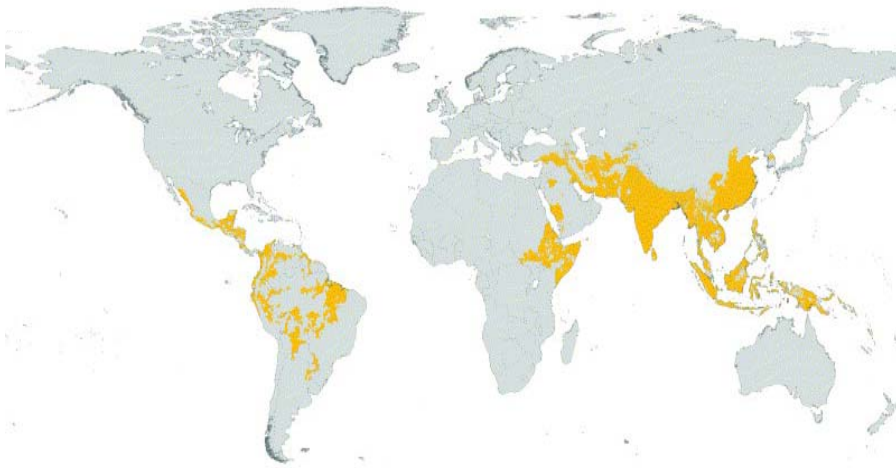
Peruvian patient isolates reveals highly

Plasmodium vivax

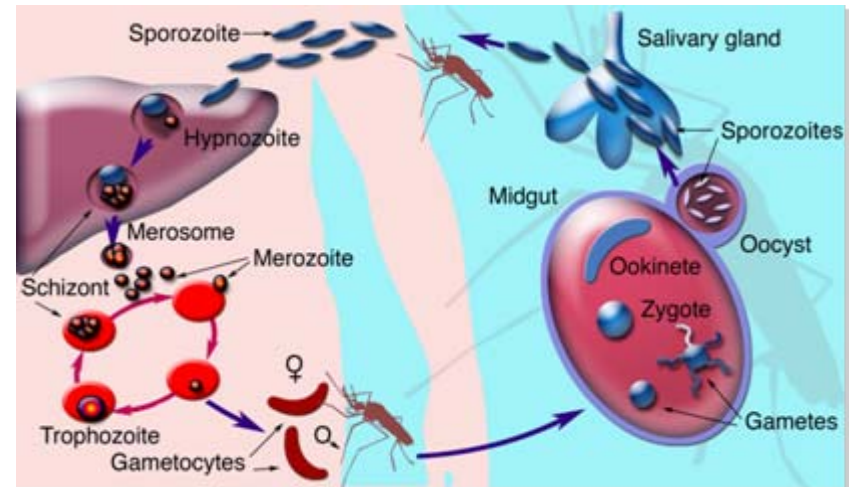


Plasmodium vivax

- Major cause of malaria outside Africa
 - 25-40% of clinical cases worldwide
- Not amenable to *in vitro* culture
- Interesting biology
 - Hypnozoites: dormant liver stage responsible for relapses
 - Primaquine: only drug for radical cure
 - Unknown determinants of chloroquine resistance



(Guerra, 2006)



(Winzeler, 2008)



P. vivax microarray analysis

- *P. vivax* infected patient blood samples from Iquitos in Peruvian Amazon, filtered to remove human white blood cells and gametocytes
- DNA was extracted, amplified and hybridized to a custom whole genome *P. vivax* tiling array



P. vivax whole genome tiling array

- Custom tiling array for *P. vivax*
 - >4 million probes of ~25 nucleotides each
 - Six base pair spacing based on sequence of Salvador I strain
 - Probes alternating sense and antisense strands
 - 1.8 million probes corresponding to 5305 *P. vivax* genes

```
TTGAGCAATTTGTACAACGTGGTGA
TTAAACATGTTGCACCACTACGCGG
TACAACGTGGTGATGCGCCTGAATG
CACCACTACGCGGACTTACGCCTCA
ATGCGCCTGAATGCGGAGTGCATTT
GACTTACGCCTCACGTAAAACCTCAA
GCGGAGTGCATTTTGAGTTGCTACT
```

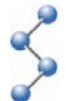
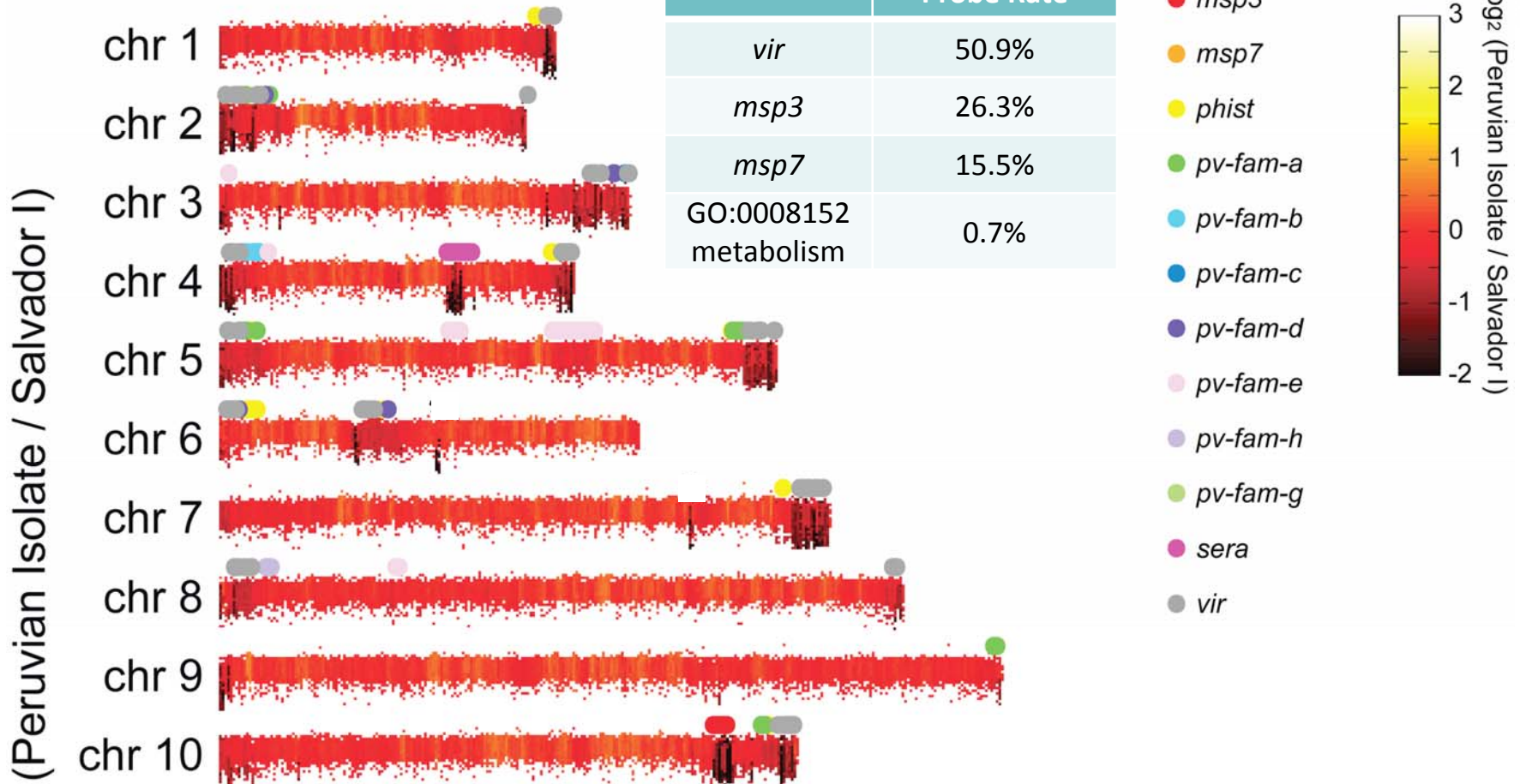
- Similar *P. falciparum* tiling array detects >90% SNPs genome-wide

(Dharia, et al. 2009, Genome Biology)



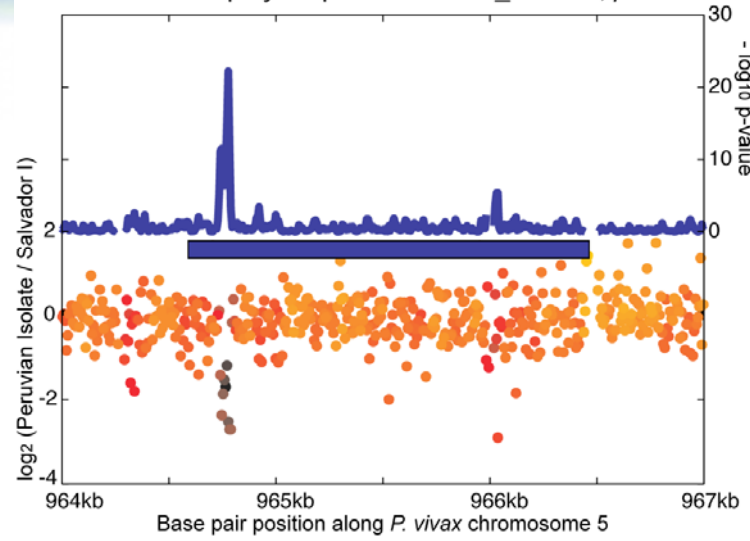
P. vivax isolate Genetic Diversity

Isolate PQSJ67

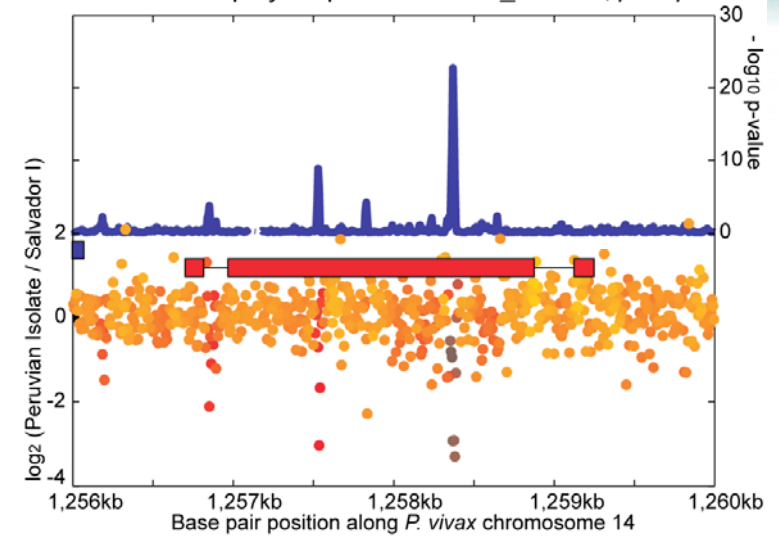


P. vivax Drug Resistance Genes

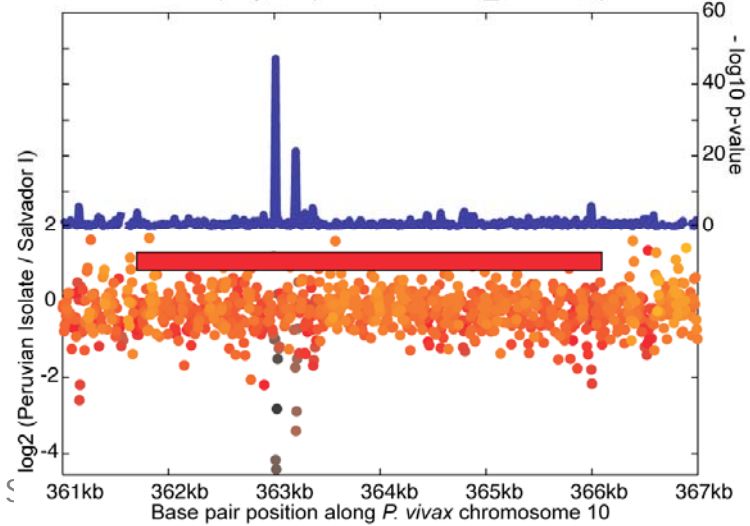
A Detection of polymorphisms in PVX_089950, *pvdhfr*



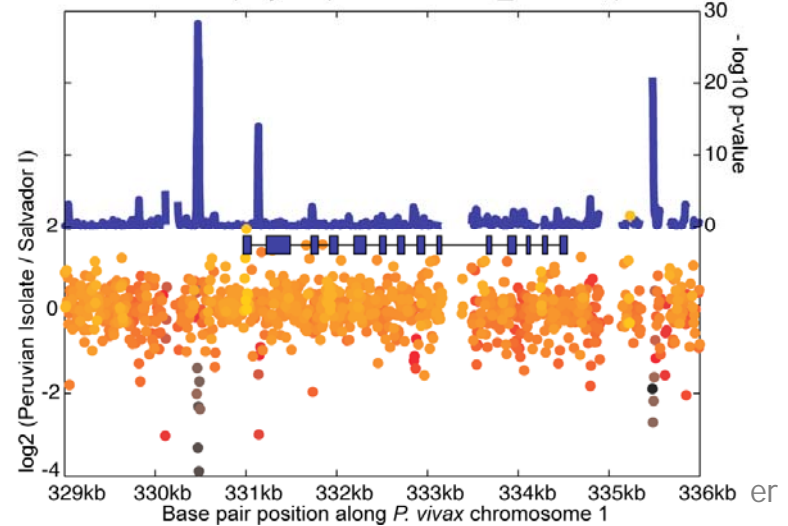
B Detection of polymorphisms in PVX_123230, *pvdhps*



C Detection of polymorphisms in PVX_080100, *pvmr1*



D Detection of polymorphisms in PVX_087980, *pvcrt*



PQSJ67 Solexa Sequencing

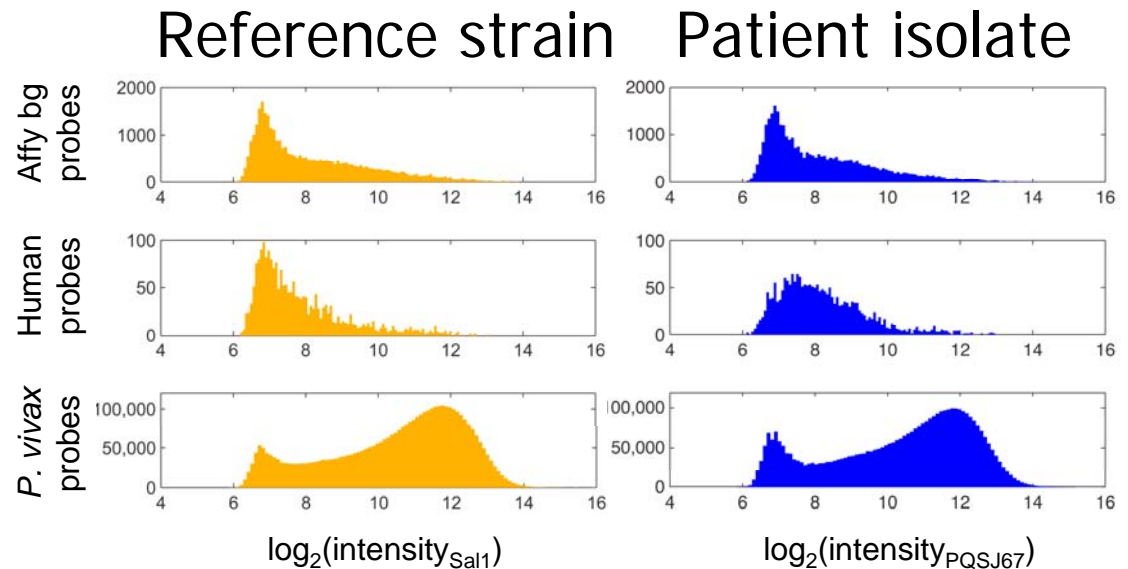
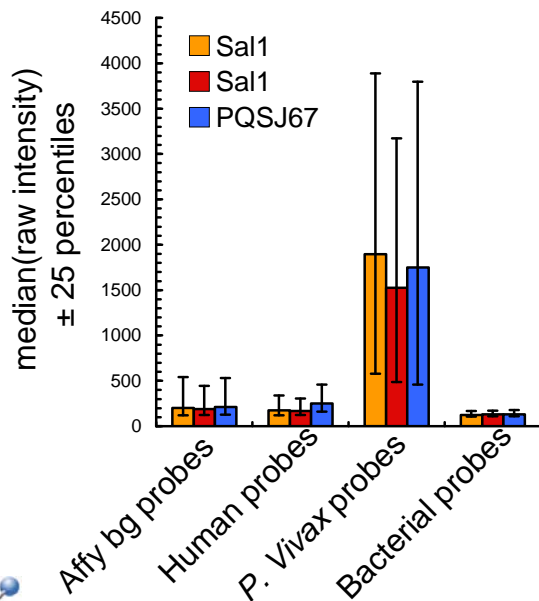
- 3 lanes of Illumina Genome Analyzer sequencing of Whole genome amplified patient sample PQSJ67
 - 28,958,672 forward; 28,958,672 reverse; 57,917,344 total reads
- Aligned reads to reference genome sequence using Bowtie software (Langmead *et al*, 2009)

Genome		Reads	% of Total Reads
<i>P. vivax</i> Genome PlasmoDB-5.5	total	23,182,796	40.03%
	unique	22,427,291	38.72%
Human Genome	total	26,530,335	45.81%
	unique	25,774,830	44.50%
Unaligned – Variable <i>P. vivax</i> sequences?		8,959,718	15.47%

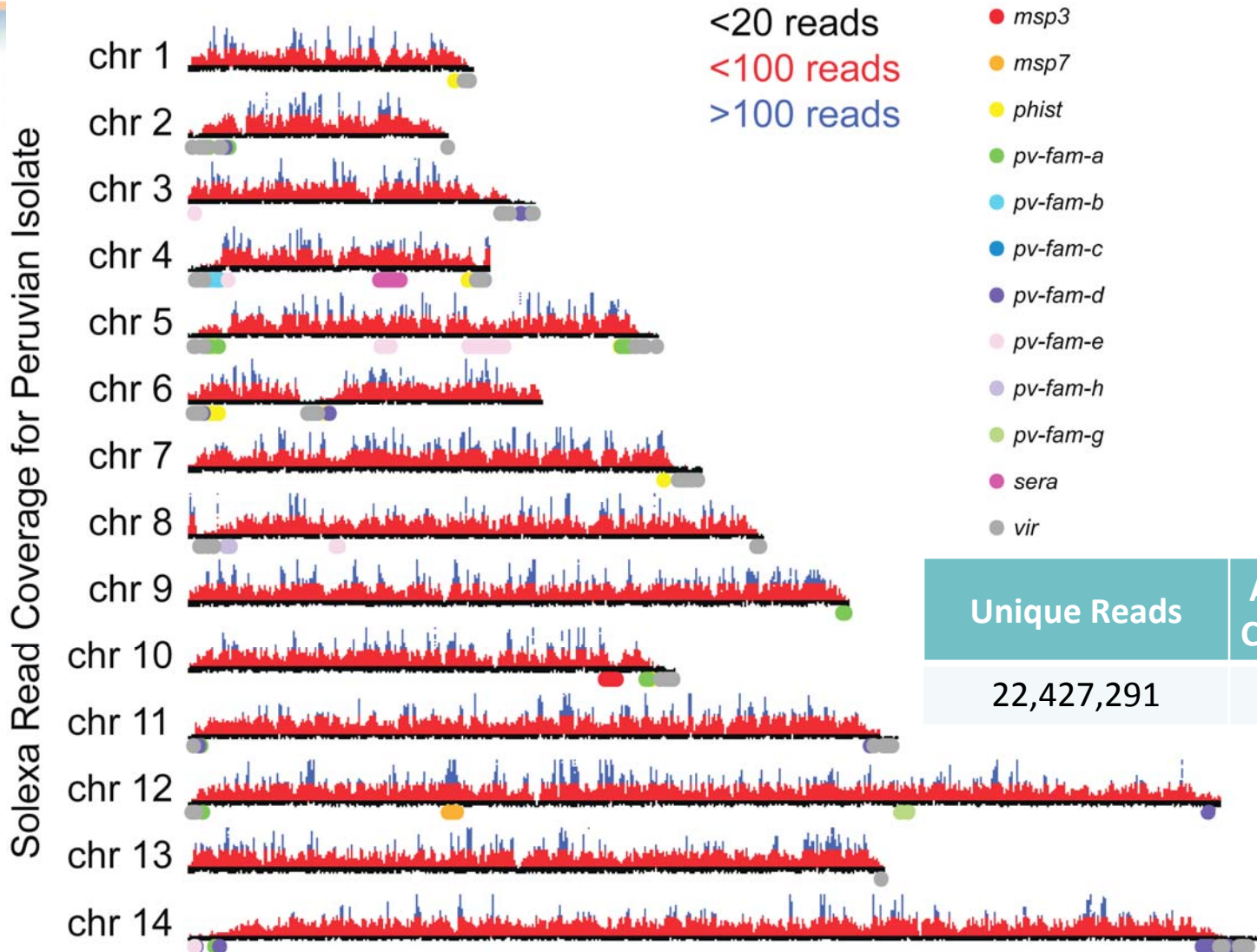


Human and Parasite DNA

Cell	DNA Content of Cell	Reads per Mb	Average Coverage	Relative Number of Cells	Relative Gene Copy Number
Human nucleated cell	6.4 Gb (diploid)	4,027	0.16	1	1
<i>P. vivax</i> parasite	26.8 Mb (haploid)	836,839	33.47	208	104
Human erythrocytes with 0.1% parasitemia	none except in parasitized cells			207,791	



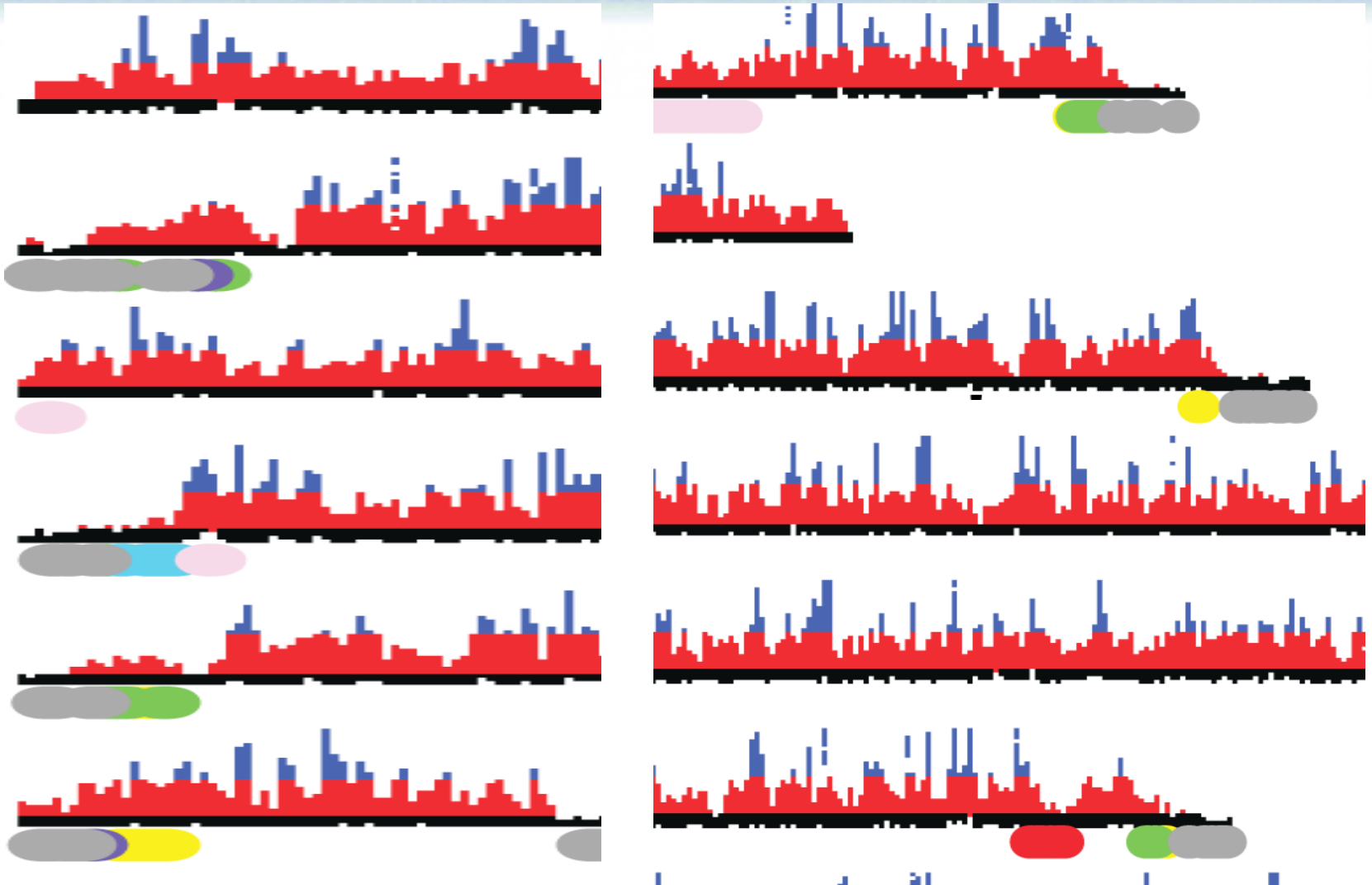
Pileup Alignment of *P. vivax* Reads



Unique Reads	Average Coverage
22,427,291	33.47

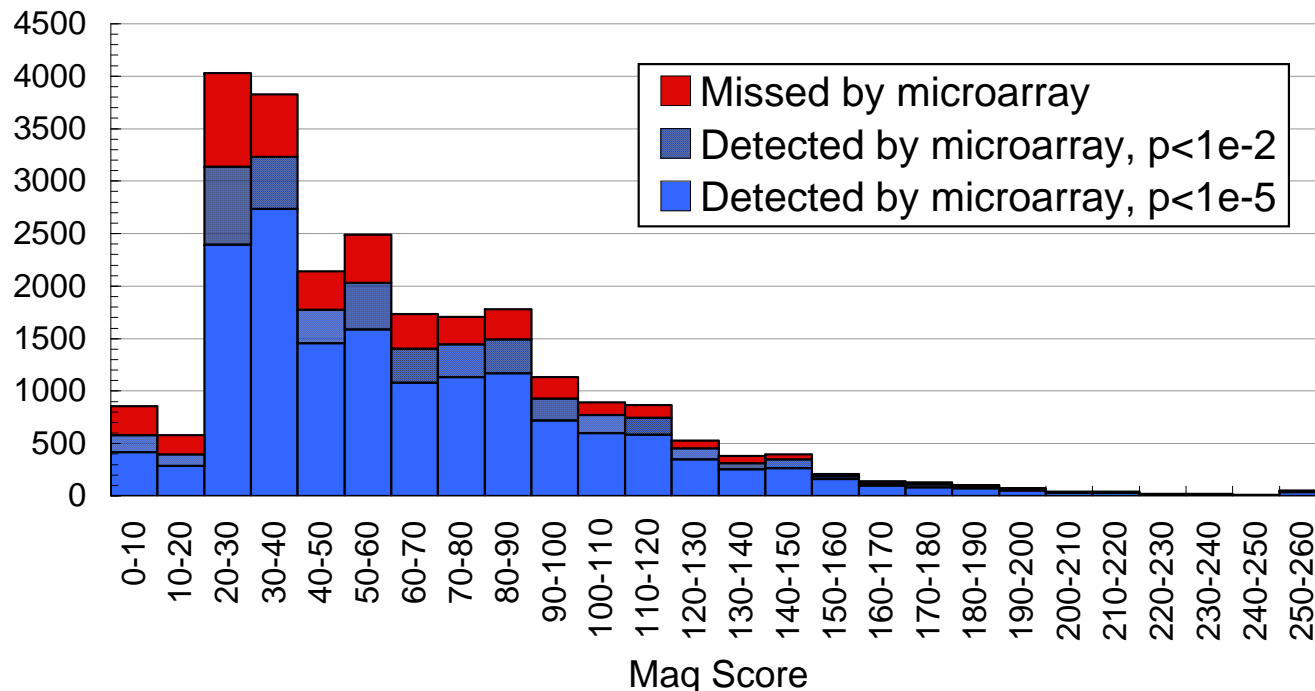


Low coverage of variable genes



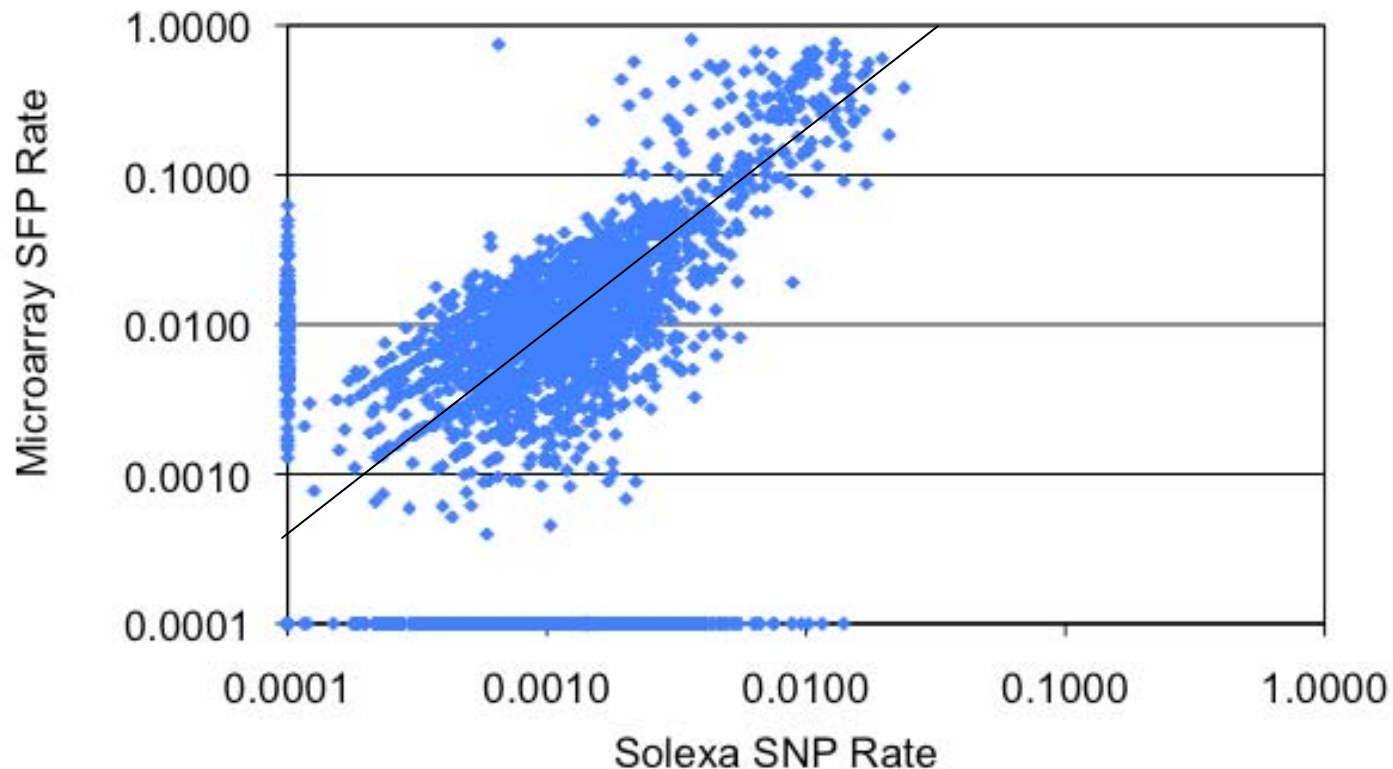
Comparison of SNP Calls

- Used Maq software (Li, 2008) to find SNPs in Solexa alignment
- SNP Quality scores based on base call and read mapping
 - 69,816 total calls
 - 41,641 Mixed SNPs, Mostly sequencing errors
 - 28,175 unmixed calls, 24,181 covered by 3 or more unique probes
 - We detect over 82% of all Solexa SNPs at $p\text{-value} < 1 \times 10^{-2}$, 65% at $< 1 \times 10^{-5}$



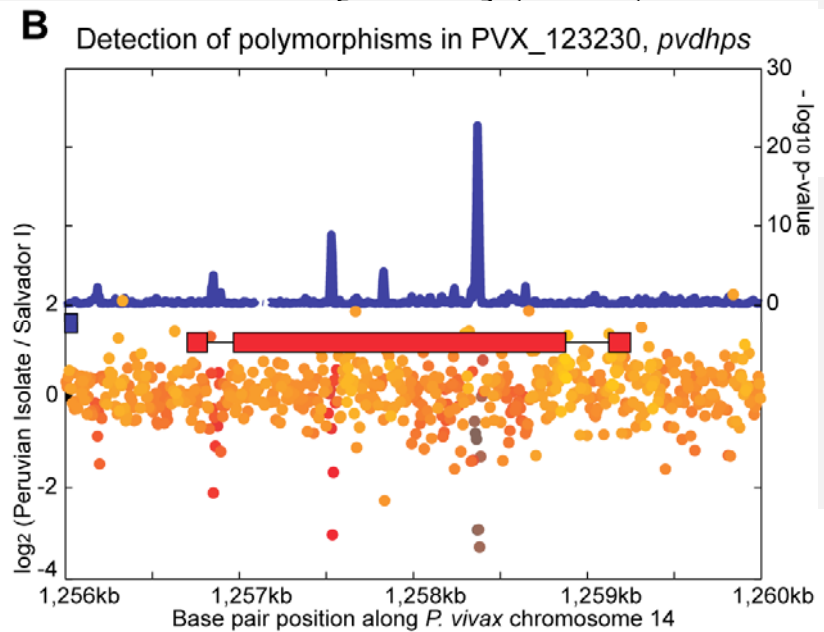
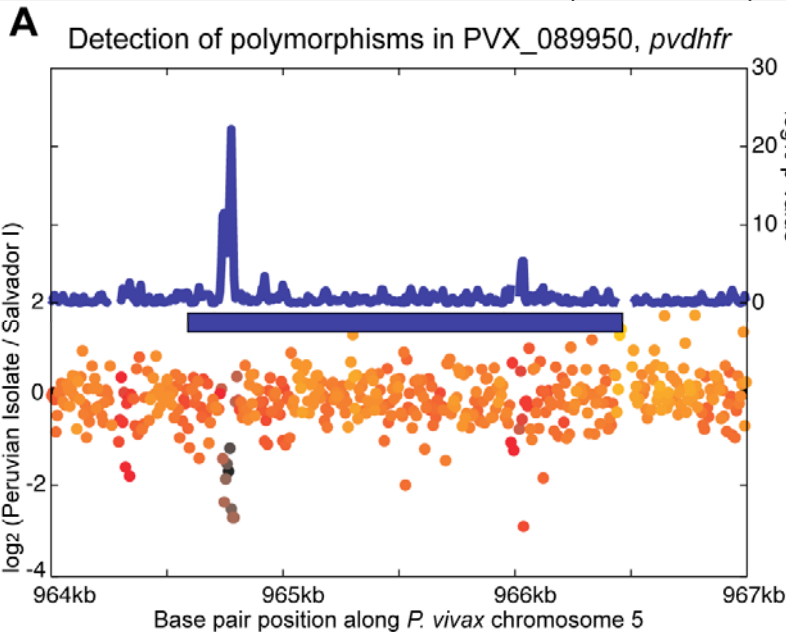
Gene SNP Rates and SFP Rates

- Positive Correlation between the two technologies
- Higher SFP rates, since one SNP will overlap multiple microarray probes, resulting in multiple SFPs



Confirmed SNPs in resistance genes

Gene	Microarray polymorphic range (prediction)	Solexa sequencing (SNP reads/total reads)
<i>pvdhfr</i>	964736-964820 (964760, 964797)	C964763G [S58R] (31/31)
	964736-964820 (964760, 964797)	T964796C [Y69Y] (13/13)
	Detected at p-value = 0.0003	G964939A [S117N] (22/23)
<i>pvdhps</i>	Detected at p-value = 6.8×10^{-5}	G1257856C [A383G] (20/20)
	1258358-1258418 (1258389)	C1258389T [M205I] (12/12)



Conclusions

- New technologies for *P. vivax* will help us generate a more complete picture of genetic diversity.
- Whole genome data allows us to study to geographical parasite populations at an unprecedented level of detail.
- Highly variable genes under positive selection by the host immune system as potential vaccine candidate antigens.
- Will dramatically accelerate drug development and discovery of resistance mechanisms.
- Loss of heterozygosity, limited diversity, and linkage disequilibrium will indicate regions under selection in drug resistant samples.
- New SNP genotyping assays can be used for tracking the spread of drug resistance in populations.

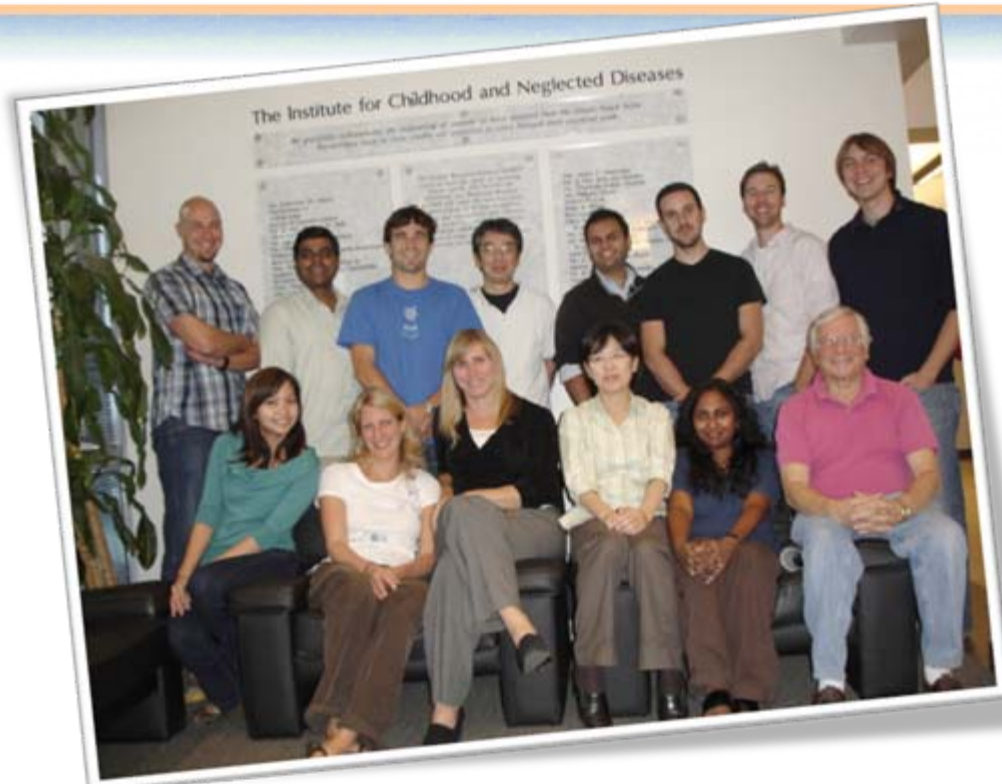


Microarray vs Sequencing

- **Microarray advantages**
 - Faster, 2 days for sample prep and all data analysis
 - Reproducible, consistent, easy data analysis
 - Cheaper per array, but greater cost investment to design and purchase the arrays in large quantities
- **Microarray disadvantages**
 - Don't know the exact polymorphism, only differences from the reference genome sequence
 - Have to re-sequence polymorphisms in regions of interest
- **Sequencing advantages**
 - Identify the exact position and base pair change
 - Can identify new sequences, with de novo assembly
- **Sequencing disadvantages**
 - Data analysis is difficult and time consuming with free software
 - Not user friendly, need bioinformatics expertise
 - Difficult to identify indels, highly variable regions have no coverage



Acknowledgements



Elizabeth A. Winzeler – The Scripps Research Inst.

- Neekesh V. Dharia
- Yingyao Zhou (GNF)
- Selina E.R. Bopp
- Stephan Meister
- Shailendra K. Sharma
- A. Taylor Bright
- Gonzalo E. González-Páez



Joseph M. Vinetz - UCSD

- Colleen McClean (Peru)
- Raul Chuquiyauri (Peru/UCSD)
- Anonymous malaria patients of Iquitos, Peru

Naval Medical Research Center Detachment (Peru)

- David J. Bacon
- Paul Graf

Centers for Disease Control

- John Barnwell
- William Collins
- Steve Hoffman (Sanaria, Inc.)

