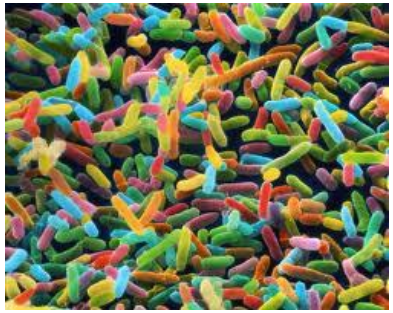
Genomics of Pathogens: Positive selection and disruptive technology

Neil Hall

@neilhall_uk

Hall Lab

- Parasites
- Microbiomes
- Plants



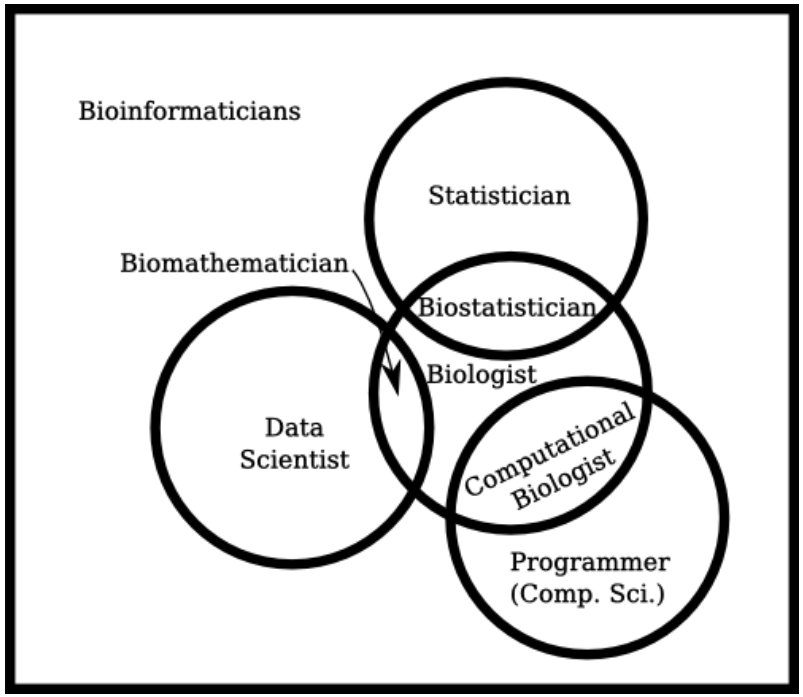
The Centre for Genomic Research

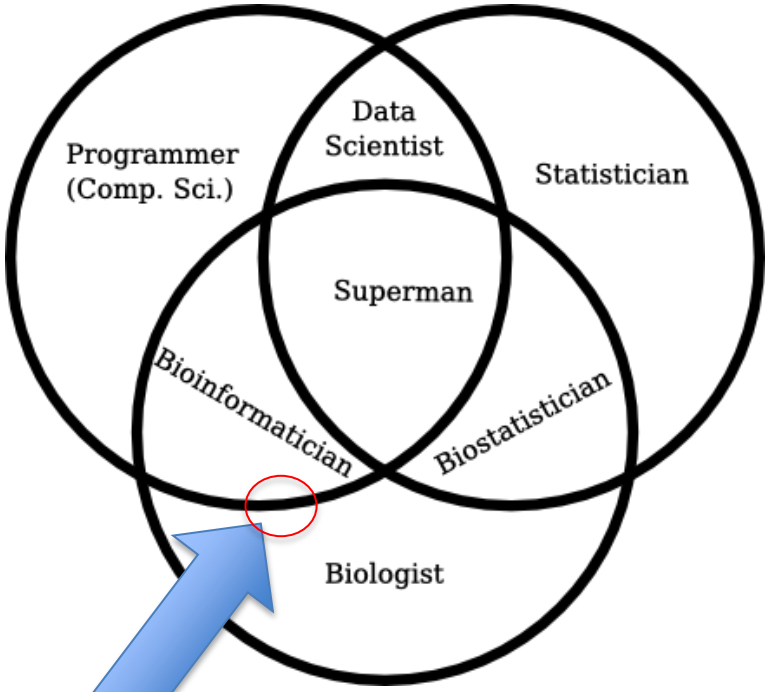
A national centre for genomic technology
Funded by
- NERC (Core funding)
- MRC (core funding)
- BBSRC

http://www.liv.ac.uk/genomic-research/

What's a bioinformatician?

(and what am I?)





Me

Grasp of Computational subjects:

More ⟷ Less

Bioinformatician ⟷ Computational Biologist

Anthony P. Fejes

http://blog.fejes.ca/?p=2418

Overview (Part 1 Pre-NGS)

- NGS
  - Why cost matters
  - NGS as an assay
- Sequencing as an assay
  - The red queen
- Genomics with 1 genome (old school genomics)
  - *Plasmodium falciparum* genome project
  - The disease and the parasite
  - The genome
  - Metabolism

Overview (Part 2)

- Genomics with >1 genome
  - What we learn from other Plasmodium spp
  - Transcriptomes, proteomes and gene regulation

- Genomics with >100 genomes
  - Population genomics in Plasmodium
  - Mapping drug resistance

DRUGS!

- Biology-by-sequencing in *Entamoeba*
  - The Entamoeba genome
  - Sex and entamoeba
  - Mapping virulence
  - Transcriptomics and more sex

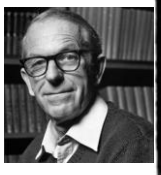
SEX!

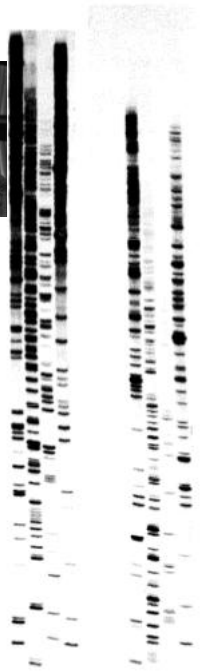
Genomics After NGS
New Technology

Genomic Evolution

1975

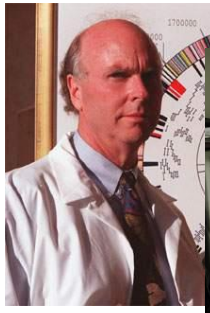
Genes

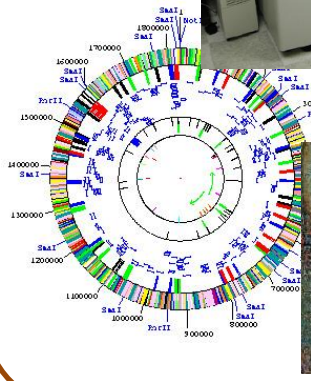




1995

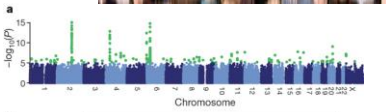
Genomes

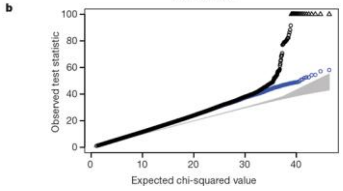




> 2007

Populations





Properties of sequence reads

```
@SRR001666.1
CACAAAGTGGCAAAGTACAATTAGTCTATAATGAACATGTTGAACTGATAGAGGTACCAATAAAACCTAGT
GATCGTTTAAAAGCTCGTGATATGTTGGGTAAAC
+SRR001666.1
BABBBFF@DFAFGGFDGGBGFGFHGHHHHHHHHHHHHHHHHHHFHHFHDHHHHFHGHHHHHHHHHHHH
GHHGHHHHHGFHGHHFHHHHHGFHFHGHHHHGHGGHFH
```

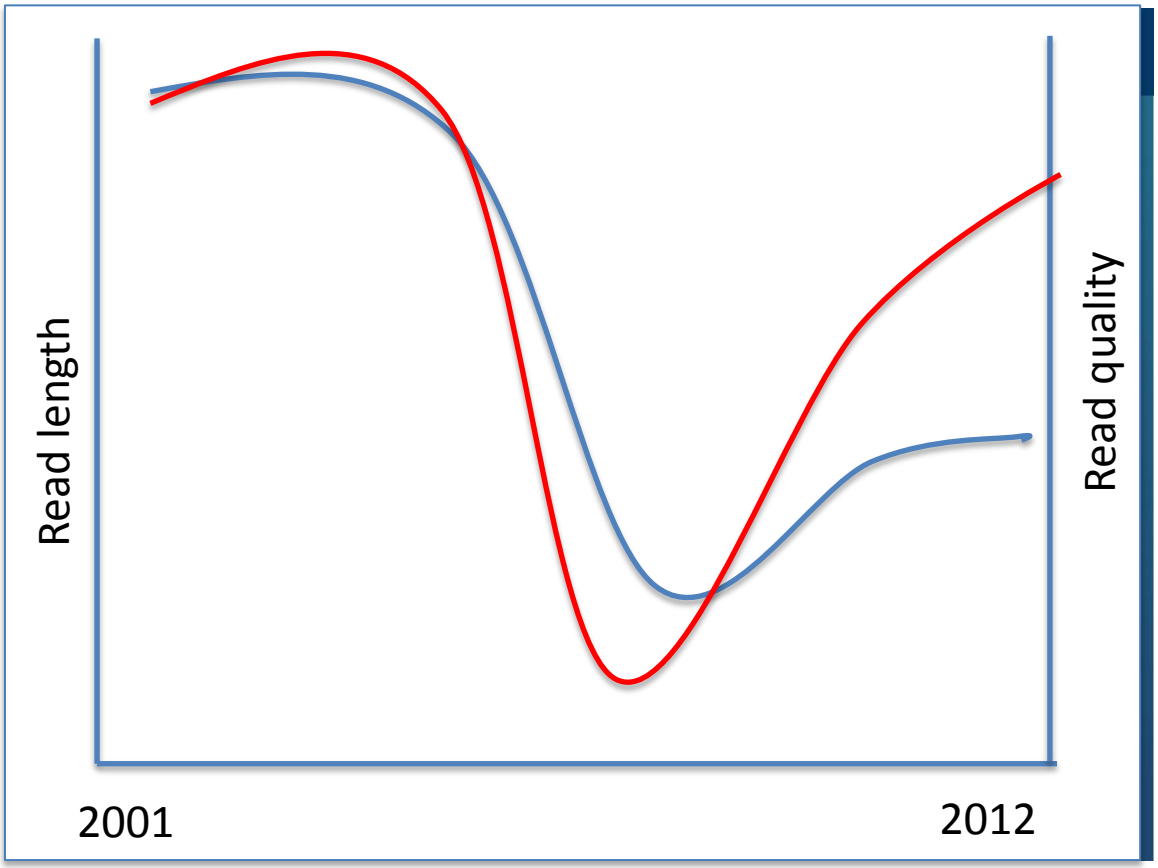
- Sequencing reads properties
  - Length
  - Quality

Qualities that scientists would say they are interested in

  - Cost
  - Time to production

Qualities that scientists are actually interested in

Revolution in costs and throughput



http://www.genome.gov/sequencingcosts/

Hall, N *Genome Biology* 2013, **14**:115

Sequencing costs have long been seen as a crucial issue for scientists

Pinning Down Sequencing Costs

How much does it cost to sequence a single base pair of DNA? The question lies at the heart of the controversy over the Human Genome Project, whose ultimate goal is to sequence all 3 billion base pairs of human DNA. But despite all the uproar over the project's cost, it turns out that no one actually knows how expensive sequencing is.

Estimates range from pennies a base up to $10 or so. In their 5-year plan for the genome project, published earlier this year, the National Institutes of Health and the Department of Energy estimated the cost at $2 to $5 a base, including DNA preparation and salaries. Then genome officials at NIH and DOE tried to get a firmer fix on the number by inviting in several big sequencing groups to talk about their costs. Their conclusion? Sequencing now costs a mere $1 to $2 a base—apparently a phenomenal improvement.

Way too low, replied members of NIH's genome advisory committee. "$1 to $2 is a dangerous estimate," said genome project director James Watson. "If you use that, you'll find the cost is going up." Stanford biologist David Botstein agreed: "No one has the vaguest idea [what sequencing costs], and that should be our position."

But genome officials are determined to nail down this figure. They have recently hired a

Science 1990

CORRESPONDENCE

Reality of sequencing costs

SIR — In their interesting article on the *Caenorhabditis elegans* genome sequencing project, J. Sulston *et al*. (*Nature* **356**, 37–41; 1992) anticipate ". . . the cost, currently estimated at $1 per base with current methods. . ." of their strategy. How did they arrive at this estimate, and what is included in estimating the cost?

Let me do the following calculation for such a project in Germany. Two fluorescence sequencers (paid off over 5 years) cost about DM80,000 per year, 600 primers synthesized (length 15, DM5–8 per base) is more than DM45,000 for 120,000 base pairs (DM0.375 per bp). DNA preparations, chemicals, enzymes, gels and so on come to at least DM15,000 for 1,500 clones (DM0.12 per bp). If, say, one-third of the authors' time was spent on laboratory work such as cloning, DNA preparation, sequencing and editing (DM75,000 × 19/3) the cost for salaries would be DM475,000 per year. Rent and running cost for a laboratory of 120 people will be about DM30,000 per year. There will be additional costs for laboratory equipment, repair, laser-tube replacement, break-down time, overheads and so on, often neglected by scientists.

With such rather optimistic numbers I arrive at minimal costs of DM645,000 for 120,000 bp per year, or $3.00 per bp of final sequence. For 800,00 bp per year with full-time employment and 20 per cent overhead, the costs triple to DM2 million and will be close to $2.00 per bp.

I founded a sequencing company in 1990, in which about 600,000 bases were read for assembly during 6 months from two cosmids by using direct-blotting electrophoresis and colorimetric detection of digoxigenin. The actual costs were close to the above estimate. The costs do not include profits and the work was done by dedicated people. This is in accord with the present funding of the yeast project — requiring very high quality for the sequence data — by the European Commission with 2 ECU per bp, which is close to $3 per bp.

But it still appears to be a very long way before costs of $1 or $0.50 per base pair will be a reality. One should be very careful with such estimates, especially in a scientific paper, because they may have far-reaching consequences for biological research.

Fritz Pohl
Fakultät fur Biologie,
Universität Konstanz,
D-7750 Konstanz, 1 Germany

SULSTON REPLIES — I agree with Pohl's essential point that estimates of DNA sequencing costs must be all inclusive and made with care. Unfortunately, in an effort to shorten the paper and in response to a referee's opinion that cost issues would not be of broad interest, we removed the detailed discussion that was in our original manuscript.

A cost of $1 per base pair of finished sequence is the current production cost, not the actual cost of sequencing the three cosmids; the latter was indeed higher because of time spent in development. We have now settled down to a production routine, and are able to estimate our actual costs reasonably confidently. In so doing, we have allowed for all the hidden extras (such as fringe benefits, supervision and replacement of minor equipment) that, as Pohl points out, are so easily overlooked, and a 50 per cent overhead for rent and running costs of the laboratory space.

The discrepancy between our estimate and that of Pohl is due to: (1) heavier use of the ABI 373A — only two machines per megabase of finished sequencer per annum; (2) in-house synthesis of oligonucleotide primers, of which only 100 are now made per cosmid; and (3) efficient use of staff by job specialization, with appropriate levels of training for each operation.

The pace of sequencing is sustainable, because the members of the groups, though certainly dedicated and flexible in their activities, are not called upon to work excessively long hours for routine production. (Development work, and writing scientific papers, is another matter.)

John Sulston
MRC Laboratory of Molecular Biology,
Hills Rd,
Cambridge CB2 2QH, UK

Law of mass action

SIR — John Maddox (*Nature* **355**, 201; 1992) scolds molecular biologists for their naive reliance on qualitative observations, and calls for a resurrection of the law of mass action to explore crucial quantitative features of molecular regulatory networks.

We agree with the doctor's diagnosis and remedy, but would like to point out that the patient's plight is not as serious as might be inferred. In addition to the "army of people called molecular biologists" who uncover the qualitative structure of molecular control mechanisms, there are squads of theoretical biologists who wield the law of mass action to explore the quantitative implications of these mechanisms. As a small selection of recent studies that illustrate the usefulness of this approach, we draw your attention to models of cyclic AMP meta-

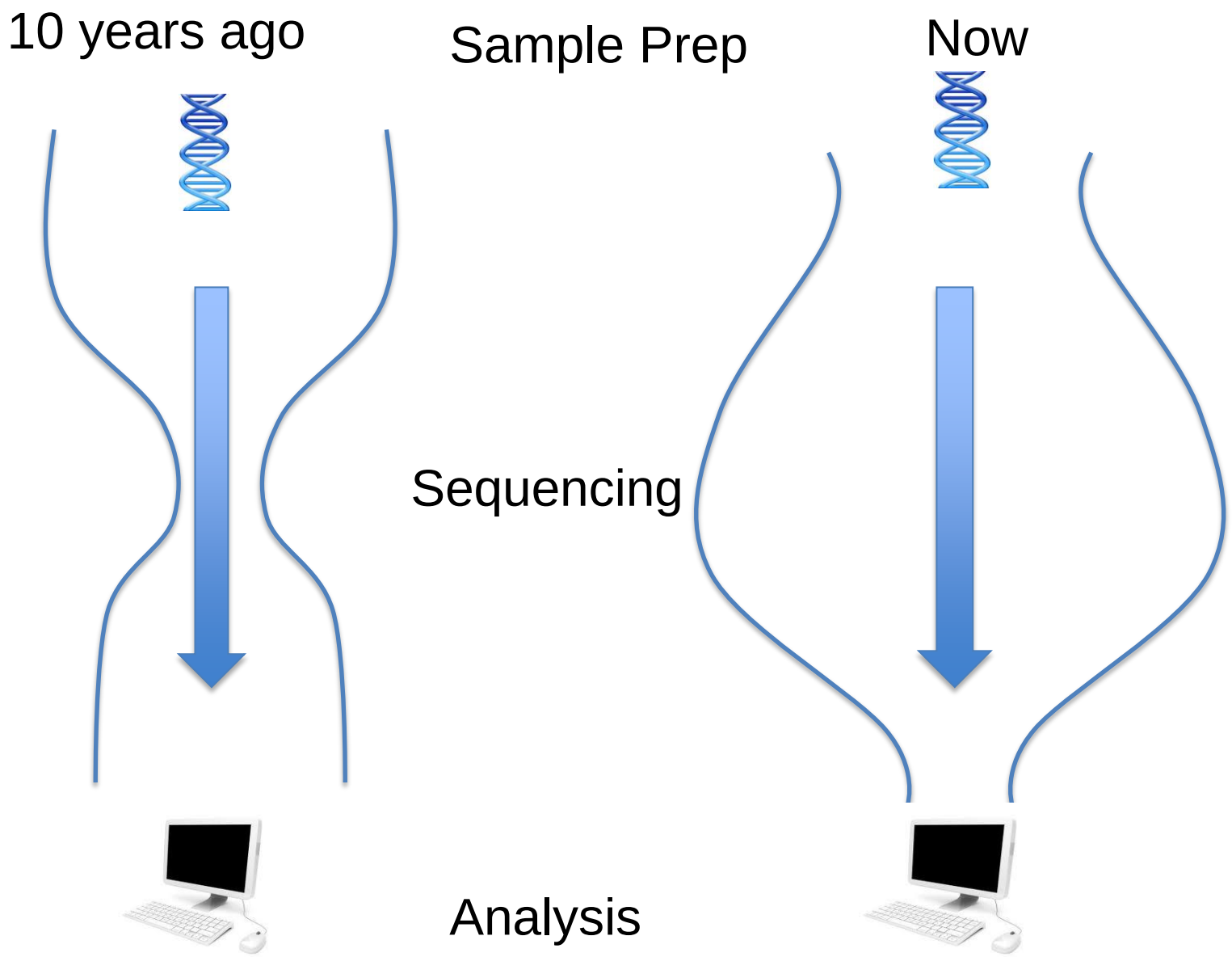
Nature 1992

Costs

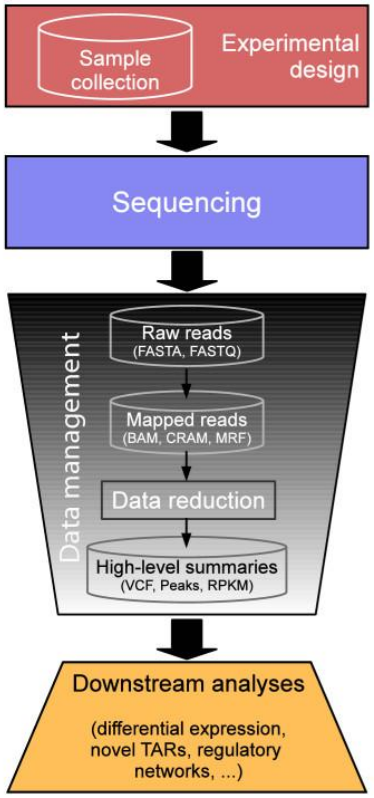
- Reagent cost
- Machine depreciation cost
- Human cost

- DNA cost

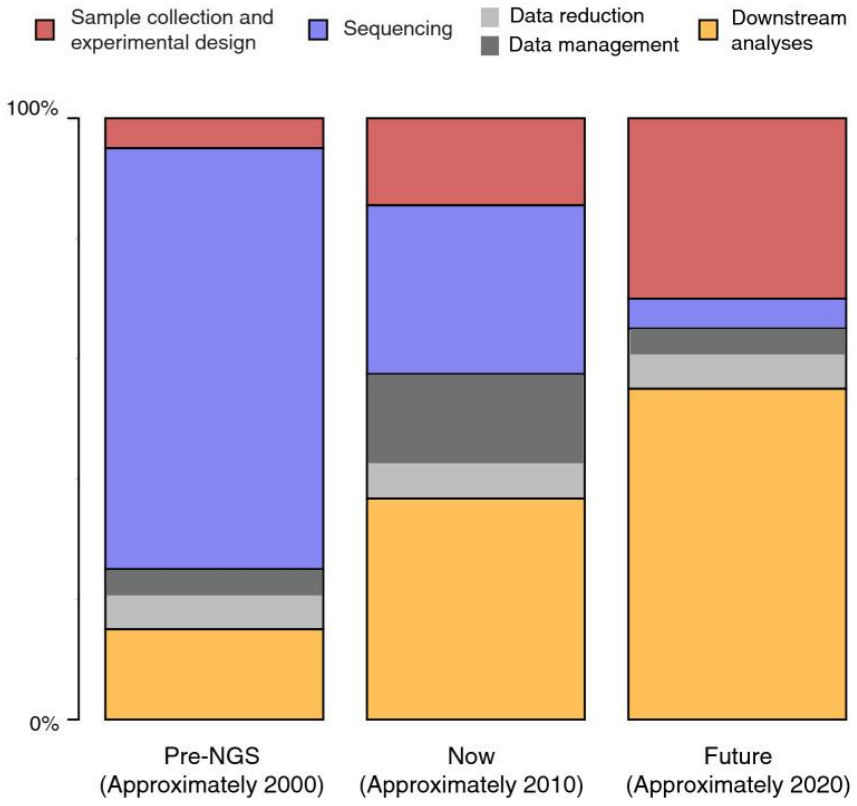
Process bottlenecks have changed



Relative costs of a genome project



Proportion of project costs



The questions we ask of the data are far more specific

Sboner et al. Genome Biology 2011 12:125

Applications

15 years ago

now

Diagnostics

Metylation profiling

Exome resequencing

Transcriptomics RNAseq

Community profiling

De novo SNP discovery Metagenomics

ChIP-Seq Mutation detection RADtags

Population genetics

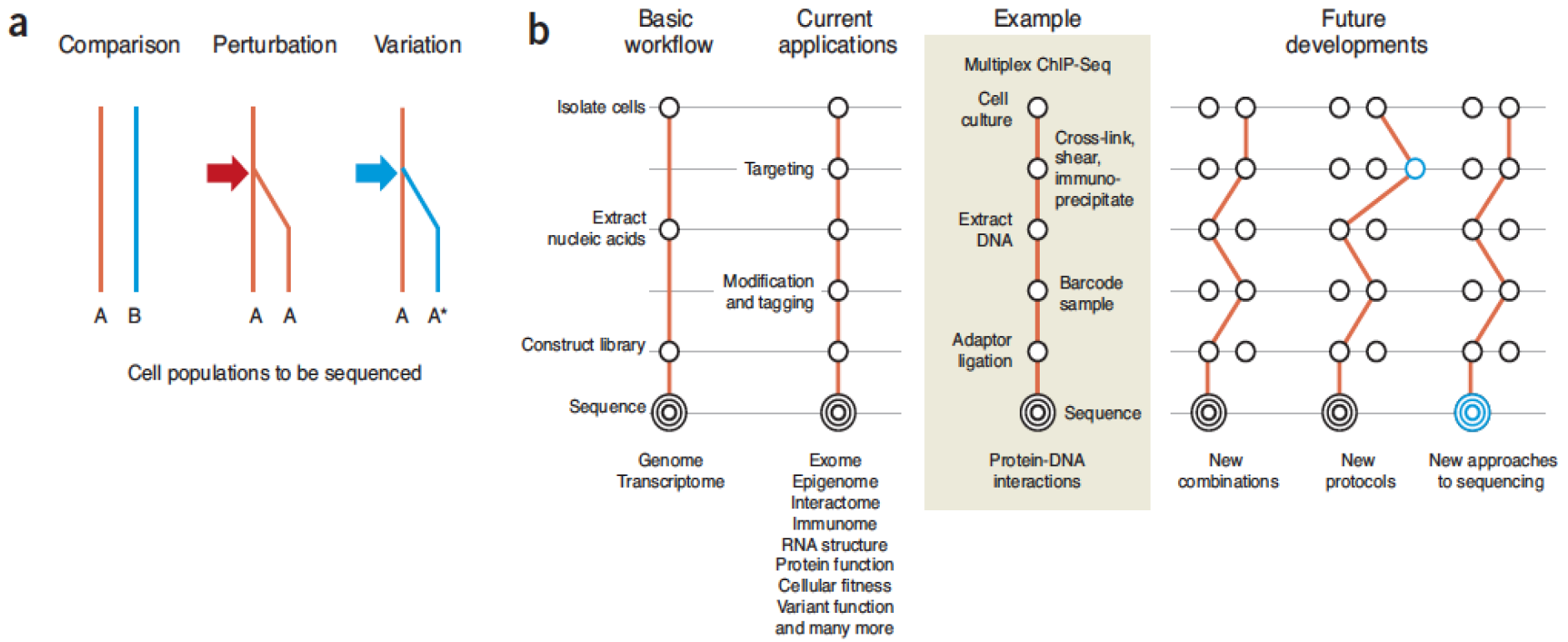
APPLICATIONS

Table 1 Applications of next-generation DNA sequencing

Method	Sequencing to determine:	Example reference	'Subway' route as defined in Figure 3
DNA-Seq	A genome sequence	57	Comparison, 'anatomic' (isolation by anatomic site), flow cytometery, DNA extraction, mechanical shearing, adaptor ligation, PCR and sequencing
Targeted DNA-Seq	A subset of a genome (for example, an exome)	20	Comparison, cell culture, DNA extraction, mechanical shearing, adaptor ligation, PCR, hybridization capture, PCR and sequencing
Methyl-Seq	Sites of DNA methylation, genome-wide	34	Perturbation, genetic manipulation, cell culture, DNA extraction, mechanical shearing, adaptor ligation, bisulfite conversion, PCR and sequencing
Targeted methyl-Seq	DNA methylation in a subset of the genome	129	Comparison, cell culture, DNA extraction, bisulfite conversion, molecular inversion probe capture, circularization, PCR and sequencing
DNase-Seq, Sono-Seq and FAIRE-Seq	Active regulatory chromatin (that is, nucleosome-depleted)	113	Perturbation, cell culture, nucleus extraction, DNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
MAINE-Seq	Histone-bound DNA (nucleosome positioning)	130	Comparison, cell culture, MNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
ChIP-Seq	Protein-DNA interactions (using chromatin immunoprecipitation)	131	Comparison, 'anatomic', cell culture, cross-linking, mechanical shearing, immunoprecipitation, DNA extraction, adaptor ligation, PCR and sequencing
RIP-Seq, CLIP-Seq, HITS-CLIP	Protein-RNA interactions	46	Variation, cross-linking, 'anatomic', RNase digestion, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, PCR and sequencing
RNA-Seq	RNA (that is, the transcriptome)	39	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
FRT-Seq	Amplification-free, strand-specific transcriptome sequencing	119	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, adaptor ligation, reverse transcription and sequencing
NET-Seq	Nascent transcription	41	Perturbation, genetic manipulation, cell culture, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, circularization, PCR and sequencing
Hi-C	Three-dimensional genome structure	71	Comparison, cell culture, cross-linking, proximity ligation, mechanical shearing, affinity purification, adaptor ligation, PCR and sequencing
Chia-PET	Long-range interactions mediated by a protein	73	Perturbation, cell culture, cross-linking, mechanical shearing, immunoprecipitation, proximity ligation, affinity purification, adaptor ligation, PCR and sequencing
Ribo-Seq	Ribosome-protected mRNA fragments (that is, active translation)	48	Comparison, cell culture, RNase digestion, ribosome purification, RNA extraction, adaptor ligation, reverse transcription, rRNA depletion, circularization, PCR and sequencing
TRAP	Genetically targeted purification of polysomal mRNAs	132	Comparison, genetic manipulation, 'anatomic', cross-linking, affinity purification, RNA extraction, poly(A) selection, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
PARS	Parallel analysis of RNA structure	42	Comparison, cell culture, RNA extraction, poly(A) selection, RNase digestion, chemical fragmentation, adaptor ligation, reverse transcription, PCR and sequencing
Synthetic saturation mutagenesis	Functional consequences of genetic variation	93	Variation, genetic manipulation, barcoding, RNA extraction, reverse transcription, PCR and sequencing
Immuno-Seq	The B-cell and T-cell repertoires	86	Perturbation, 'anatomic', DNA extraction, PCR and sequencing
Deep protein mutagenesis	Protein binding activity of synthetic peptide libraries or variants	95	Variation, genetic manipulation, phage display, *in vitro* competitive binding, DNA extraction, PCR and sequencing
PhIT-Seq	Relative fitness of cells containing disruptive insertions in diverse genes	92	Variation, genetic manipulation, cell culture, competitive growth, linear amplification, adaptor ligation, PCR and sequencing

Shendure and Aiden Nature Biotechnology 30, 1084–1094 (2012)

Sequencing as experiments



Shendure and Aiden Nature Biotechnology 30, 1084–1094 (2012)

All experiments have essentially the same workflow……

The Red Queen
A hypothesis driven sequencing experiment

"Now, here, you see", said the Red Queen to Alice in Lewis Carroll's Through the Looking Glass, "it takes all the running you can do, to keep in the same place".

Vol 464|11 March 2010|doi:10.1038/nature08798 nature

LETTERS

Antagonistic coevolution accelerates molecular evolution

Steve Paterson[1]*, Tom Vogwill[1]*, Angus Buckling[2], Rebecca Benmayor[2], Andrew J. Spiers[3], Nicholas R. Thomson[4], Mike Quail[4], Frances Smith[4], Danielle Walker[4], Ben Libberton[1], Andrew Fenton[1], Neil Hall[1] & Michael A. Brockhurst[1]*

The Red Queen hypothesis proposes that coevolution of interacting species (such as hosts and parasites) should drive molecular evolution through continual natural selection for adaptation and counter-adaptation[1–3]. Although the divergence observed at some host-resistance[4–6] and parasite-infectivity[7–9] genes is consistent with this, the long time periods typically required to study coevolution have so far prevented any direct empirical test. Here we show, using experimental populations of the bacterium *Pseudomonas fluorescens* SBW25 and its viral parasite, phage Φ2 (refs 10, 11), that the rate of molecular evolution in the phage was far higher when both bacterium and phage coevolved with each other than when phage evolved against a constant host genotype. Coevolution also resulted in far greater genetic divergence between replicate populations, which was correlated with the range of hosts that coevolved phage were able to infect. Consistent with this, the most rapidly evolving phage genes under coevolution were those involved in host infection. These results demonstrate, at both the genomic and phenotypic level, that antagonistic coevolution is a cause of rapid and divergent evolution, and is likely to be a major driver of evolutionary change within species.

between this rapid phenotypic evolution and the underlying pattern of molecular evolution has not been resolved. Crucially, it is possible to separate bacteria and phage when transferring populations to fresh media[14], which allows one partner to be held evolutionarily constant while the other partner is allowed to evolve[15–17]. Initially isogenic, replicate populations of *P. fluorescens* and Φ2 were propagated by serial transfer under two conditions: (1) evolution, in which the bacterial genotype was held constant and only the phage was allowed to adapt, and (2) coevolution, in which both the bacterium and the phage were allowed to evolve adaptations and counter-adaptations. At the end of the selection experiment we obtained whole-genome sequences of phage populations by high coverage second-generation sequencing to determine the identity and frequency of mutations in each population. Mutations were partitioned into synonymous and non-synonymous changes; very few synonymous mutations were observed and only non-synonymous mutations were used in analyses (see Supplementary Information; note that each indel (that is, insertions or deletions) was counted as one mutation regardless of its length). From these data we calculated the number of sites that had acquired mutations in each population relative to the ancestral ref-

Patterson et al 2012 Nature 464:275

Haldane 1949 – Disease and Evolution

- ‘genetical diversity as regards resistance to disease is vastly greater than that regards resistance to predators’
- ‘it is much easier for a mouse to get a set of genes which enable it to resist *Bacillus typhimurium* than a set which enable it to resist cats’
  - Reprinted in ‘Evolution’ Mark Ridley

Polymorphism vs divergence

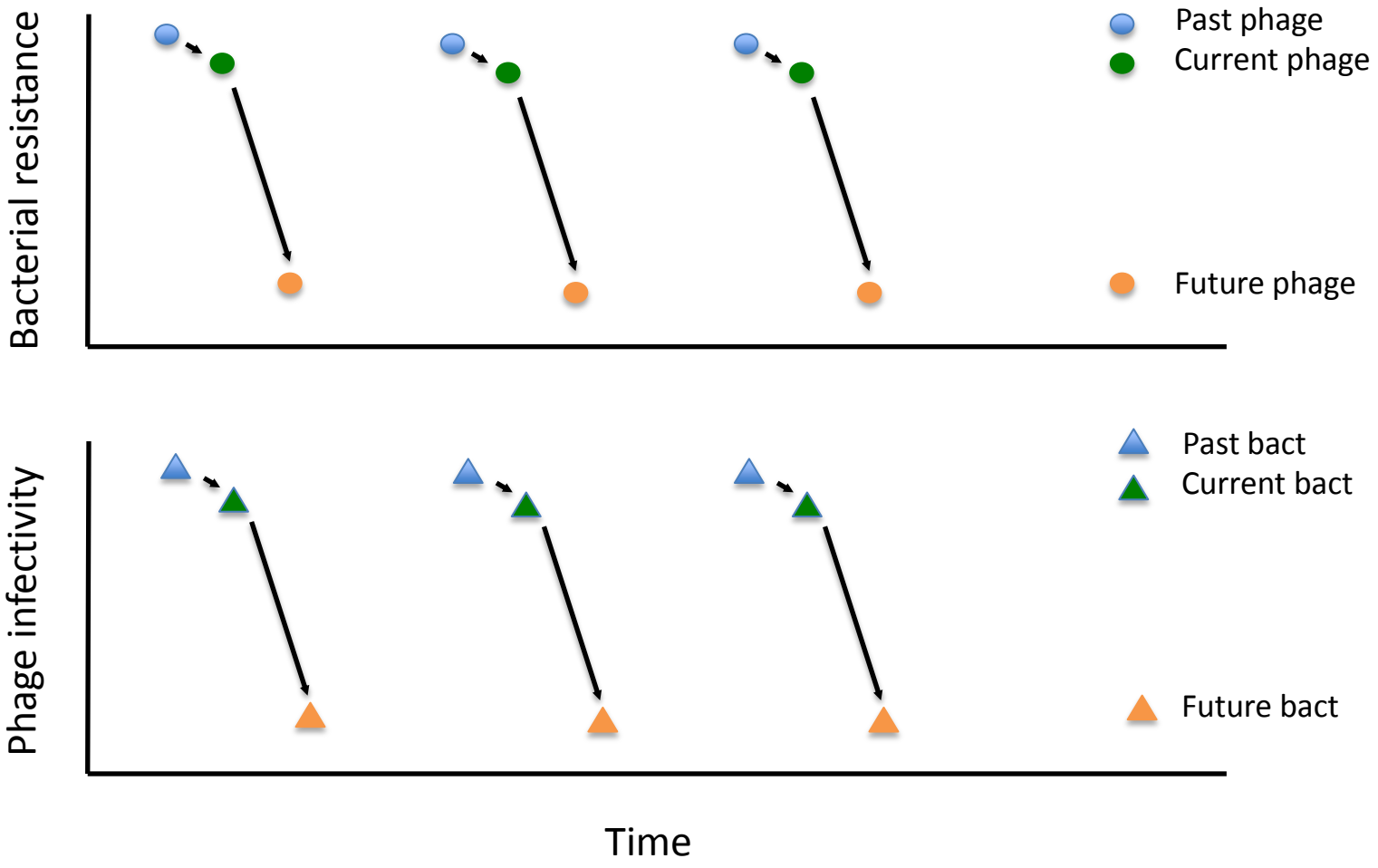
- Can use polymorphism within a species to generate null hypothesis for divergence between species

- Null hypothesis
(MacDonald-Kreitman)

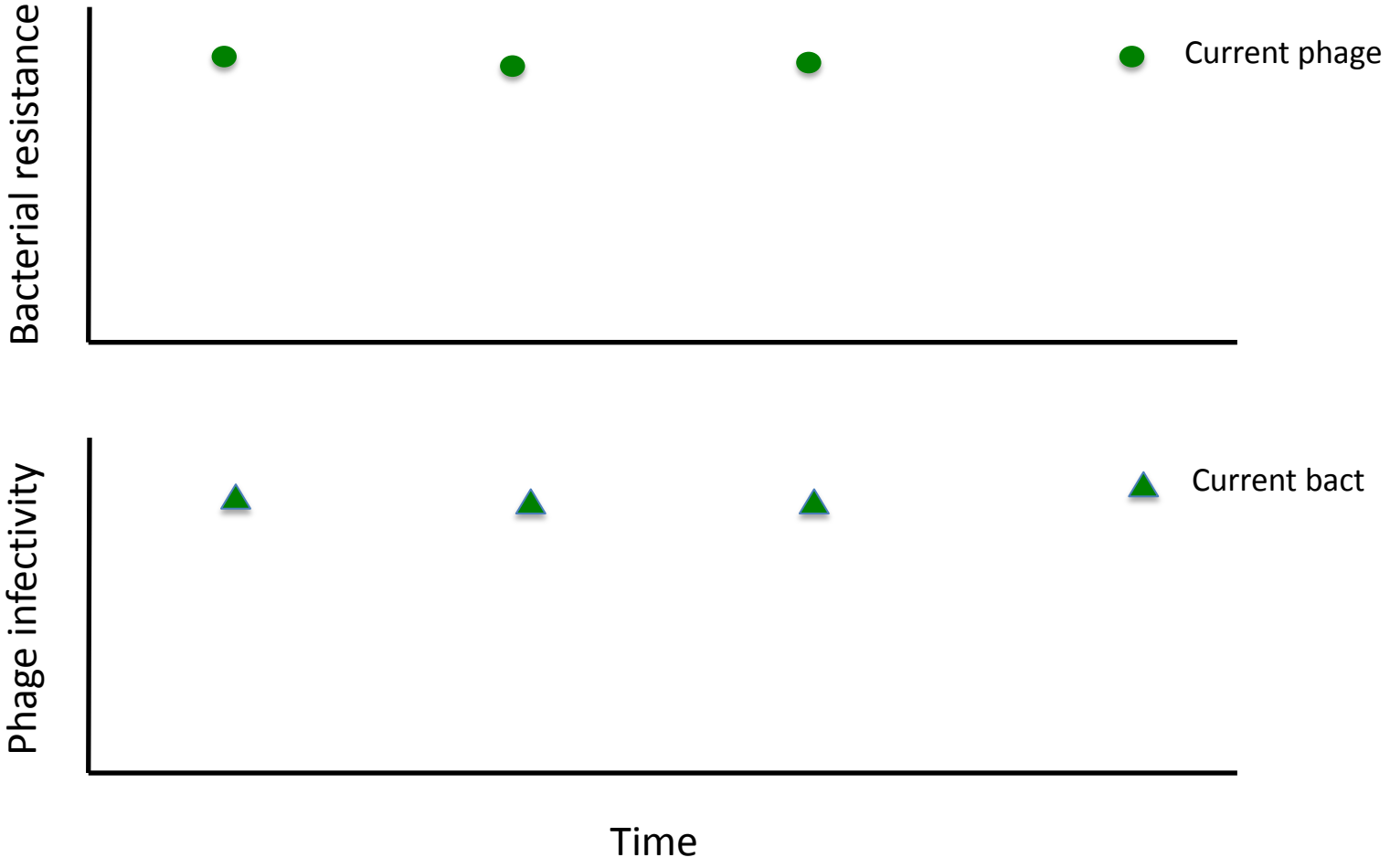
$$\frac{D_N}{D_S} = \frac{P_N}{P_S}$$

- positive selection should spread rapidly and become fixed.
  - $D_n/D_s > P_n/P_s$
- Under purifying selection
  - $D_n/D_s < P_n/P_s$

Pseudomonas & phage



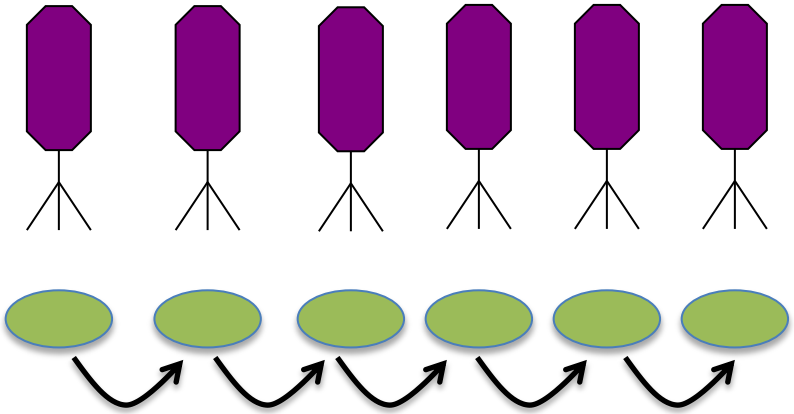
Pseudomonas & phage evolution through time



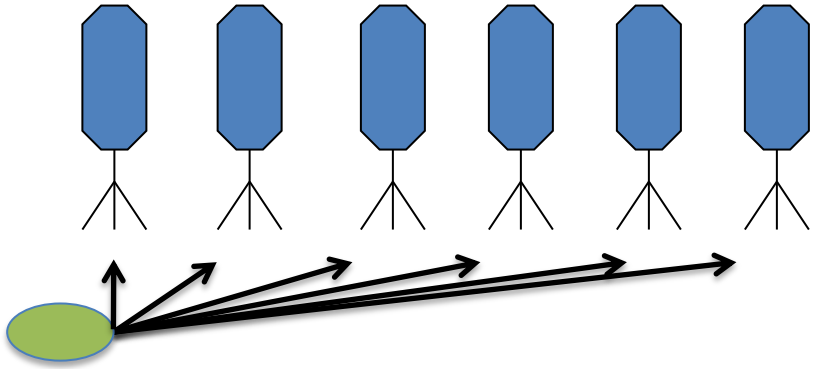
Patterson et al 2012 Nature 464:275

Experimental evolution

Φ2 phage in Pseudomonas fluorescens



Phage adapt to hosts
and hosts adapt to phage
'**co-evolution**'



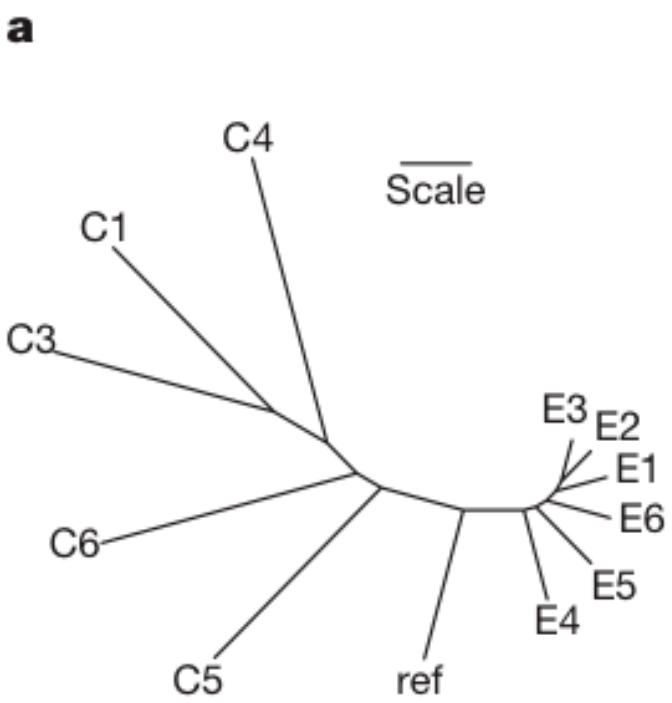
Phage adapt to fixed host
genotype '**evolution**'

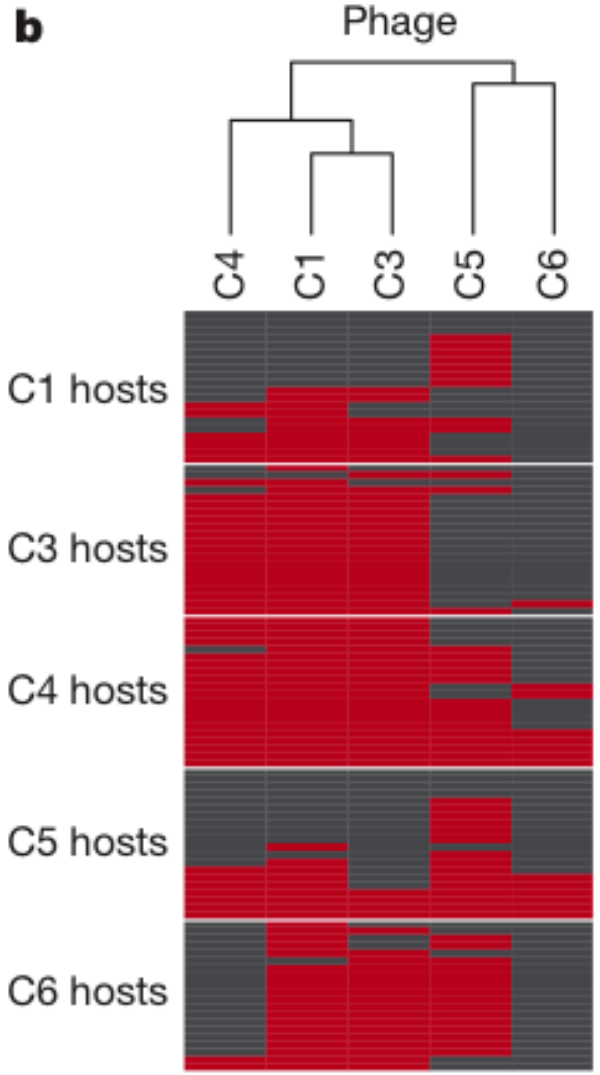
Patterson et al 2012 Nature 464:275

Next-gen sequencing

- 454 sequencing of replicate populations
  - c. 200x coverage
- Measure frequencies of variants arising by mutation and subsequently selected
  - Rate of molecular evolution with and without Red Queen?
  - Diversity within and divergence between populations?
  - What genes?

Patterson et al 2012 Nature 464:275





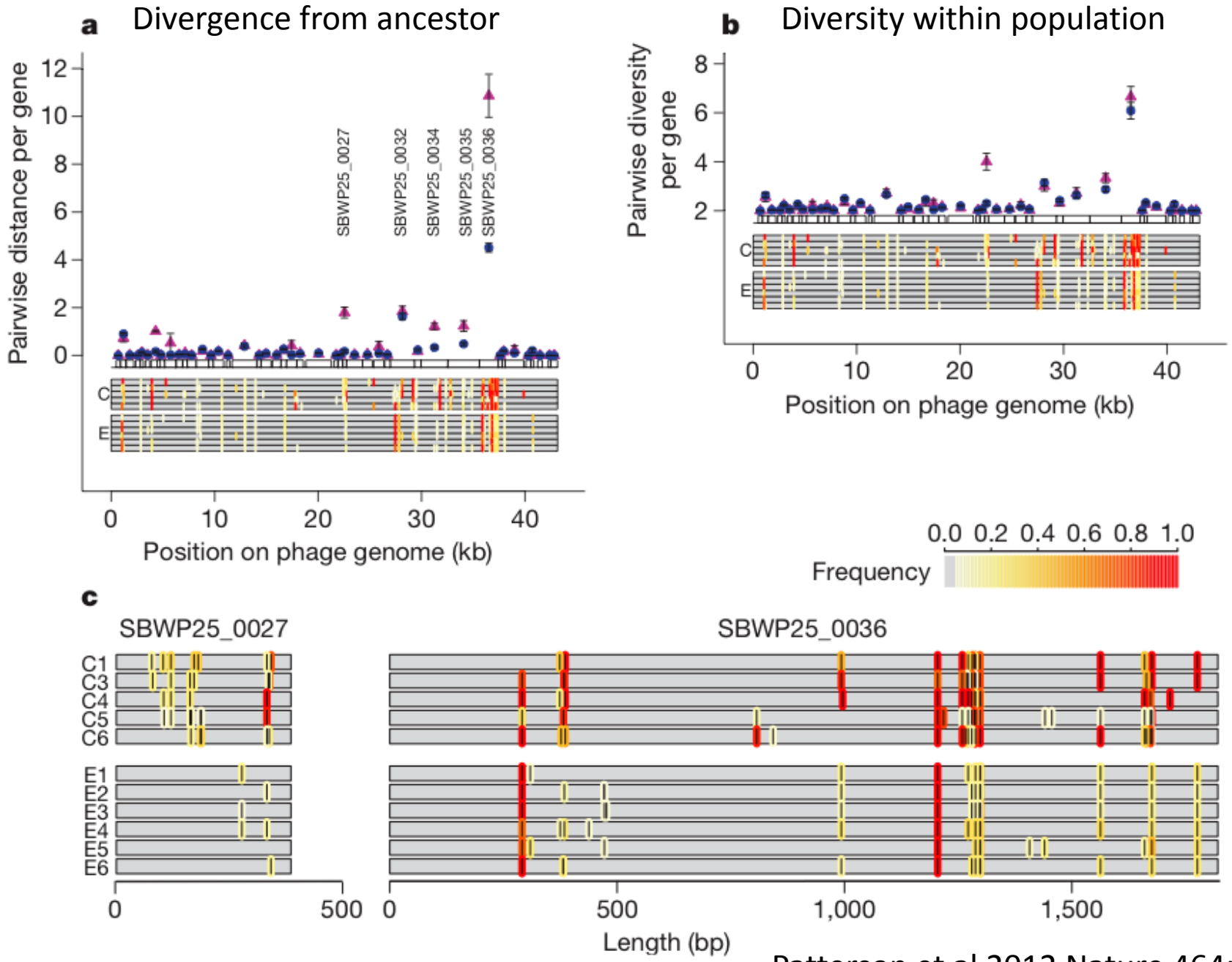
Replicates:

(1) followed a similar trajectory away from the ancestral sequence as they adapted to laboratory conditions;

(2) evolved similarly among replicates within a treatment but differently in different treatments

(3) showed independent evolution within each replicate, and higher rate in the coevolved than the evolved treatment.

Patterson et al 2012 Nature 464:275



Patterson et al 2012 Nature 464:275

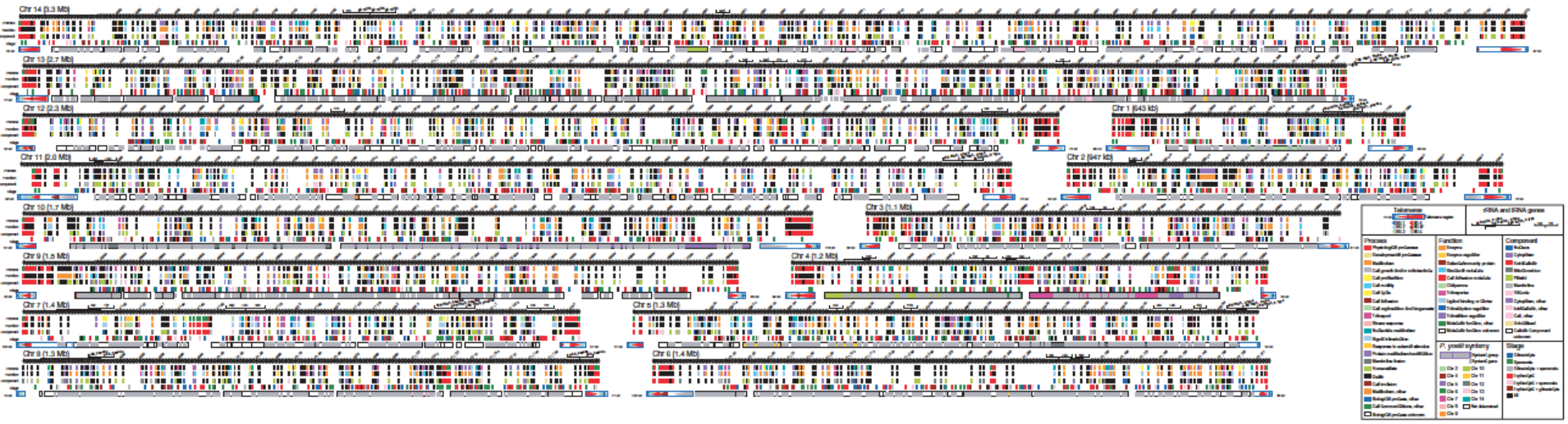
Analysis of molecular variation within and between populations

(a) Coevolved treatment only	
Within populations (σ^2_{WP})	8.49 (55%)
Among populations (σ^2_{AP})	6.89 (45%)
Total (σ^2_T)	15.37
(b) Evolved treatment only	
Within populations (σ^2_{WP})	6.65 (94%)
Among populations (σ^2_{AP})	0.42 (6%)
Total (σ^2_T)	7.07

Conclusions

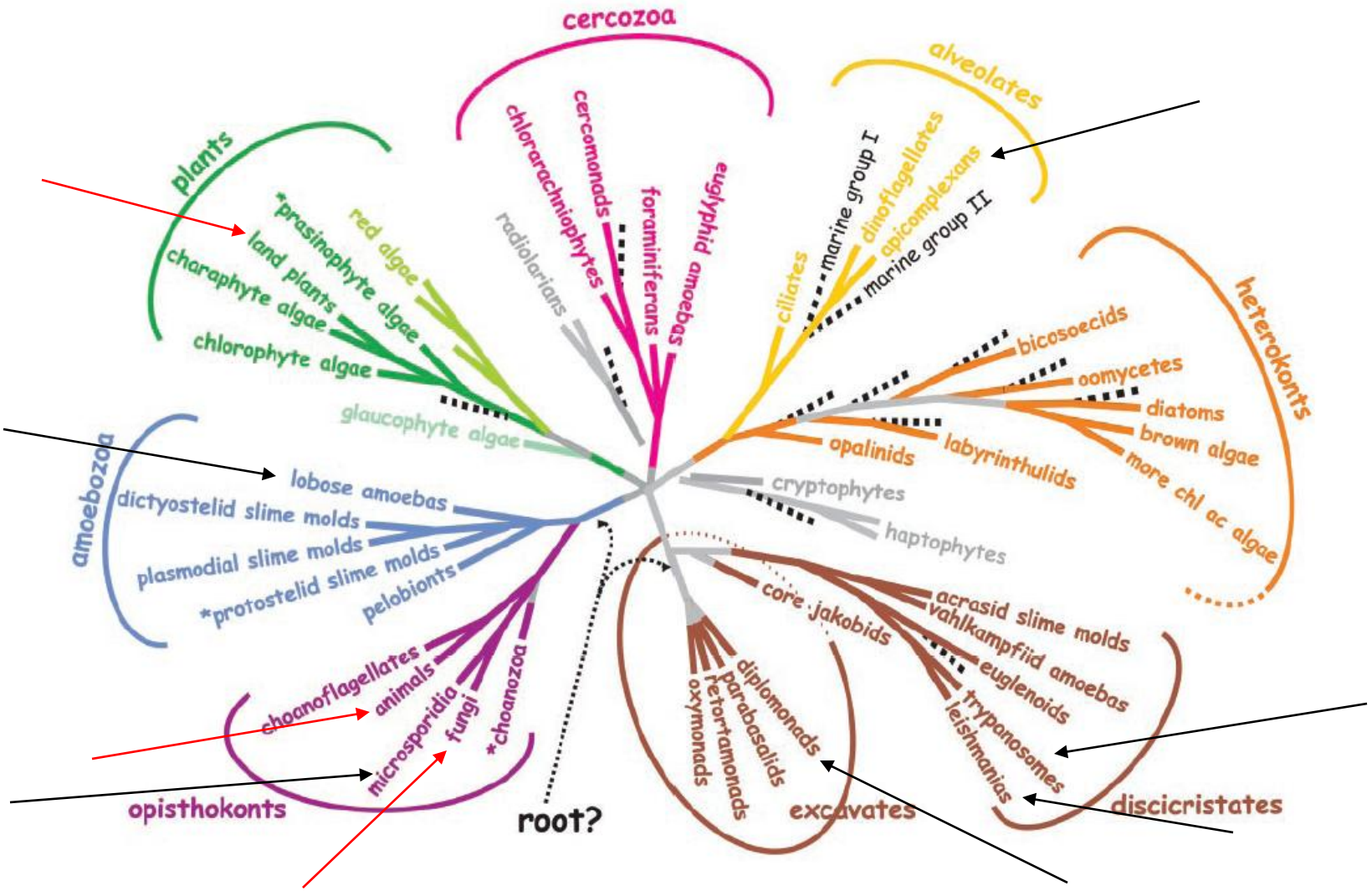
- Positive selection is stronger in the co-evolved phage.
- Co-evolved phage show twice the genetic distance from the ancestor
- Phage attachment proteins in the tail contain the most diversity within and between populations
- Suggests that there is fluctuating selection within populations

Genome of *Plasmodium falciparum*



Protist Parasite Genomes: An evolutionary perspective

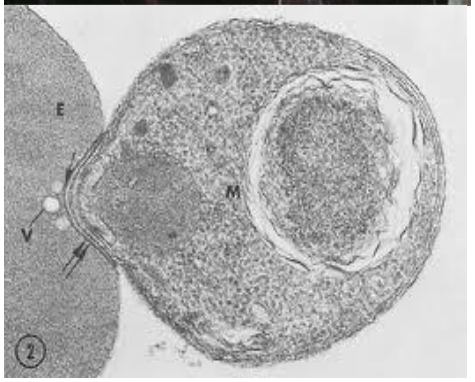
Baldauf et al (2003) Science **300:**1703



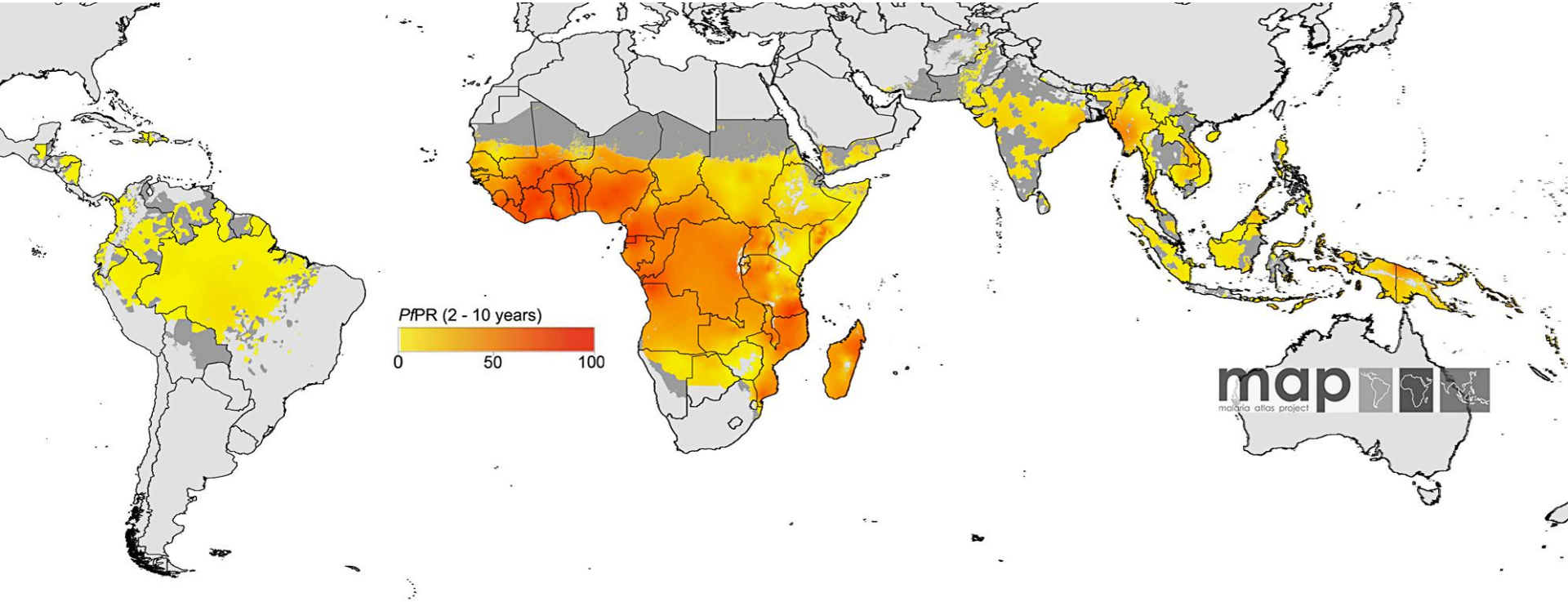
Malaria

- 300-500 million cases,1.5-2 million deaths/yr
  - Malaria kills a child every 40 seconds
- Resistance to the cheapest antimalarial drug, chloroquine, is present in almost all endemic countries
- Resistance has developed to most of the “new” antimalarials
- No practical vaccine available
- New methods for prevention and treatment of malaria are required

- Caused by *Plasmodium* spp.
- A eukaryotic single celled organism
- Complex life cycle

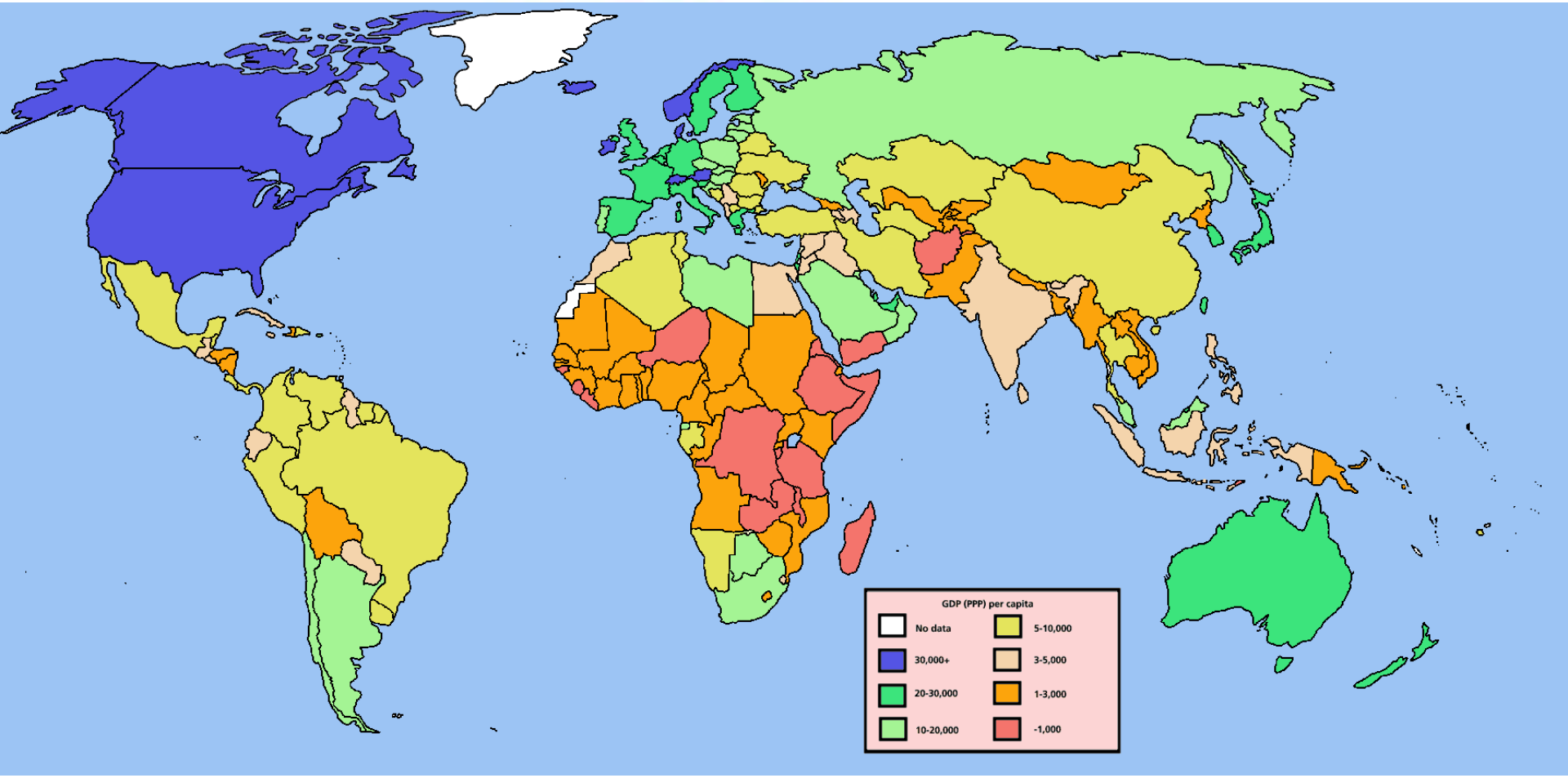


Global distribution of malaria in 2007



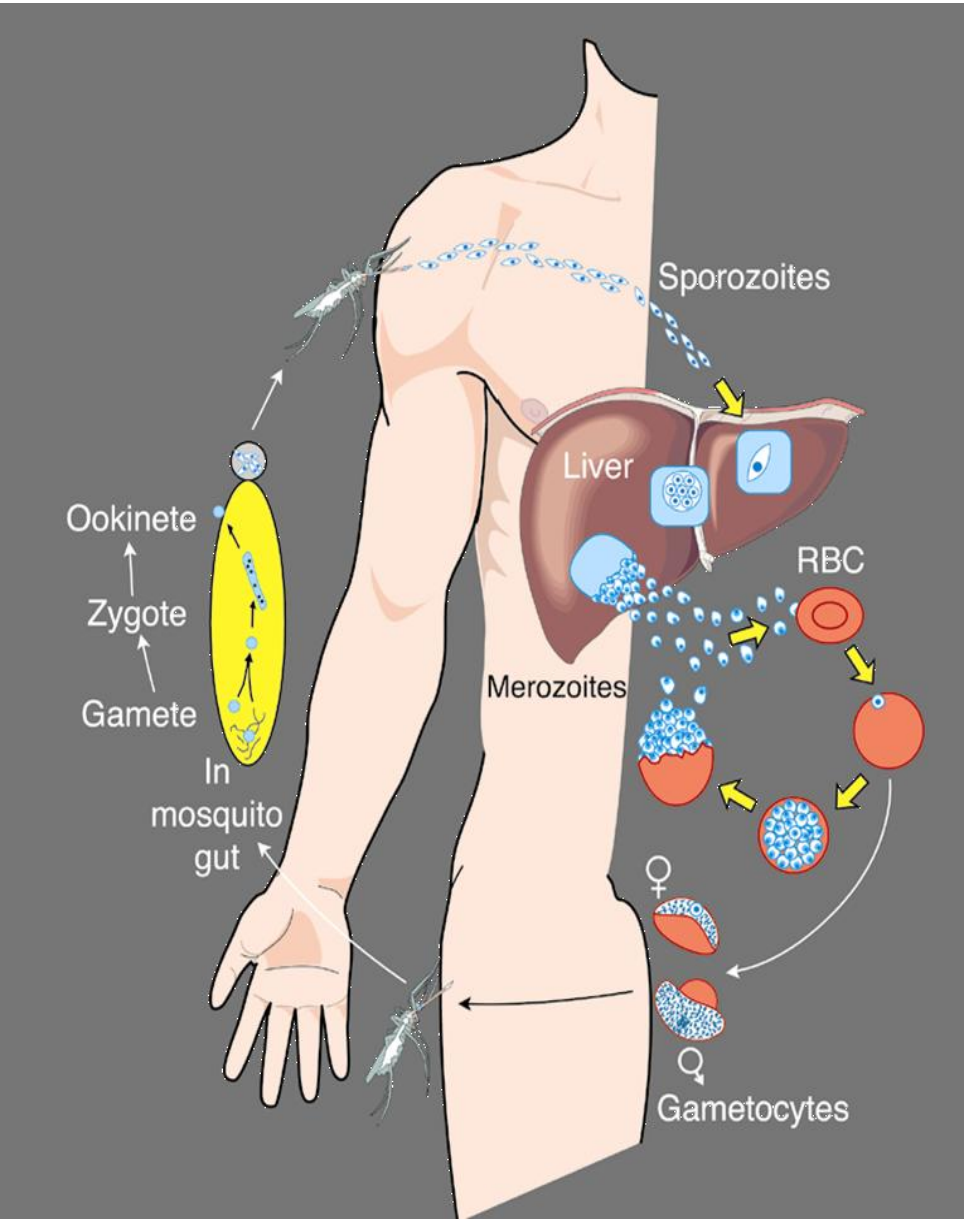
Hey et al (2009) Plos Medicine 6:e1000048 (figures 2007)

GDP per capita, 2006



Map of GDP per capita in 2006 based on IMF figures.
Map from Wikimedia.org

The Plasmodium life cycle



The malaria parasite is intracellular, spending much of its life cycle living and dividing within host cells.

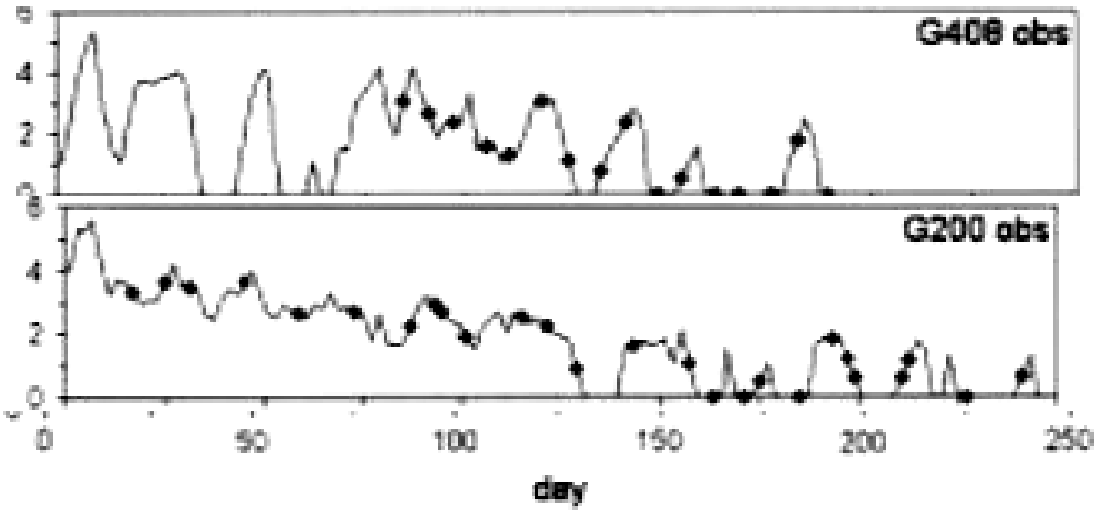
Avoids host immune system and receives a plentiful supply of nutrients.

Vaccines against malaria must target liver stages, sexual stages or invasive stages.

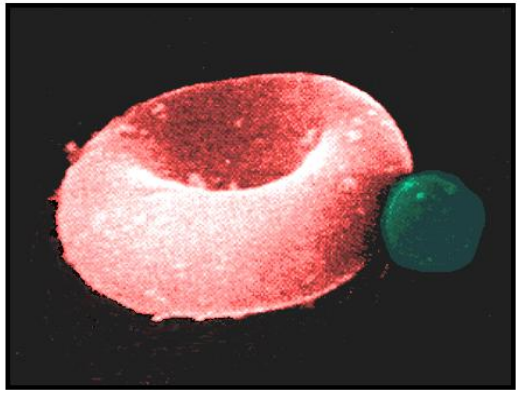
Drugs need to access the parasite in RBC or liver stage.

Infection in non-immune humans;

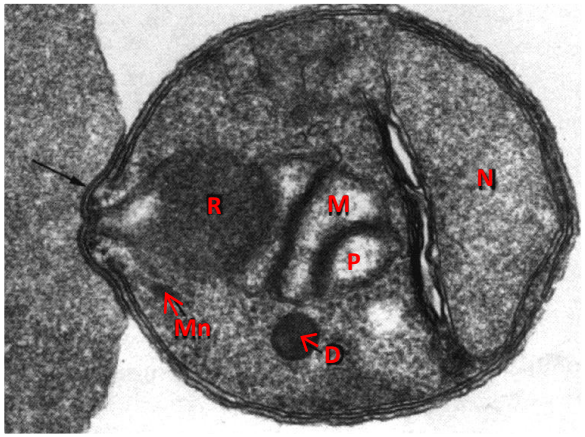
data derived from treatment of neuro-syphillis with malaria

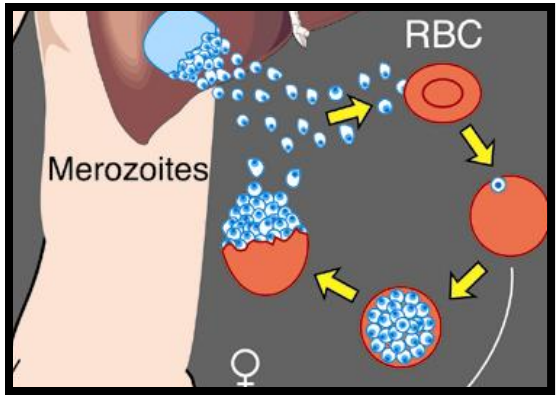


The parasite keeps changing to evade the immune response and so persists for a long time

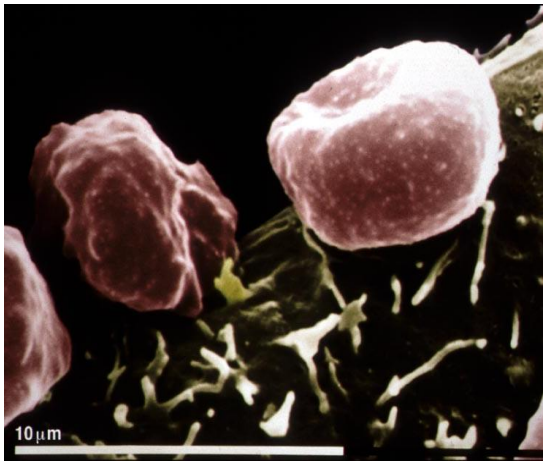


Erythrocyte invasion



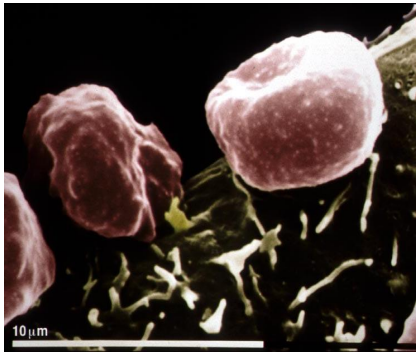


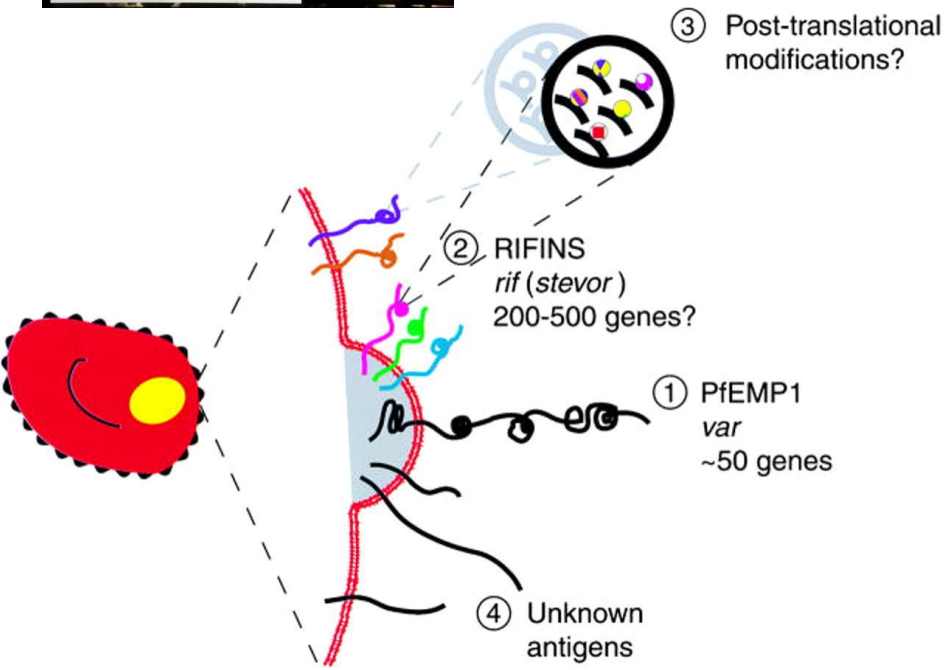
R = rhoptries
Mn = micronemes
D = dense granules



Cytoadherence and antigenic variation

Cytoadherence and antigenic variation are crucial to pathogenicity in *P. falciparum*





- The parasite expresses proteins that are exported to the red blood cell surface
- The PfEMP proteins (encoded by var genes) mediate interaction with host endothelial cells causing cytoadherence.
- Cytoadherence causes cerebral malaria that can lead to death
- Var and Rif genes are also antigenic causing immune response
- The expressed genes can change (antigeneic variation)

Fernandez et al. 193 190 (10): 1393

The Plasmodium genome project: A brief history

1995

Wellcome Trust funds chromosome 1 pilot

1996

US funds $1m pilot project

Consortium established
Sanger, TIGR , Stanford
Oxford, NMRI

1999
Chromosome 3
Published by Sanger

1998
Chromosome 2
Published By TIGR

2002 Whole genome published

Gardener, Hall, and Fung et al (2002) Nature 419:498-511
Hall et al (2002) Nature 419:527-31
Lasonder E et al (2002) 419:537-42

Not everyone celebrated

David Horrobin Enthusiasts for genomics have co rupted scientific endeavour and undermined hopes of medical progress

Not in the genes

Triumphalism about molecular biology, genomics and the human genome project is an increasingly pervasive theme in biomedical science. Beginning soon after Watson and Crick deciphered the structure of DNA, it has become progressively more dominant, reaching a crescendo over the past five years. Popular science magazines and the general media have united in a chorus of praise concerning the supposedly dramatic effects this is going to have on human health. But genomics enthusiasts have corrupted scientific endeavour and destroyed real hope of progress.

From the 1930s to the 1960s, biomedical science bore some resemblance to an integrated whole. There were researchers working at every level of biological organisation — from sub-cellular biochemistry, to whole cells, to organs, to animals, to humans. This was a golden age. Anyone receiving the best medical care in 1965 was incomparably better off than anyone in 1930.

Solutions came from many different levels. Public health specialists working at social, cultural, educational and economic levels were largely responsible for the control of malaria. Cell culture specialists working with clinicians solved polio. Microbiologists working with chemists and clinicians eliminated the threat posed by most infections. Clinicians and pharmacologists working with chemists produced the revolution in psychiatry which enabled so many patients to leave hospitals. In every field, progress depended on a constant exchange of knowledge.

But starting in the 1960s, molecular biologists and genomics specialists took over biomedical science. Everything was to be understood completely at the molecular genomic level. Everything wa to be reduced to the genome. Journals and grant-giving bodies came to be dominated by reductionists who were scathing about the complexity of whole-organ, whole-animal and especially whole-human studies which were seen as too full of uncontrolled variability to be interpretable. Clinical and physiological studies lost out and progressively their research communities were destroyed. Now we have an almost wholly reductionist biomedical community which repeatedly makes exaggerated claims about how it is going t revolutionise medical treatment — and which repeatedly fails to achieve anything.

The first genetic disease to be fully defined in molecular terms was sickle cell disease, the abnormality of the haemoglobin in human red blood cells which causes such devastation in African-origin communities. A single abnormality in a single protein causes the trouble. The abnormal protein was identified in the 1940s, the precise molecular defect was identified in the 1950s, and the three-dimensional structure of the protein was defined in the 1960s. Yet what has been the clinical impact of this wonderfully precise molecular knowledge? Precisely nothing. The clinical picture of the disease cannot yet be understood in terms of the molecular biology.

In only two areas — stomach and duodenal ulceration and organ transplantation – has medical treatment improved dramatically. In almost every other field, people are little better off than those receiving the best treatment in 1965. Take cancer. We have made dramatic progress in a very limited range of rare cancers: the leukaemias, lymphomas and testicular tumours. But even there many of the drugs used were introduced before 1965. We have simply learned to use them better.

Another example is the hype surrounding the malaria genome project. This is indeed a brilliant technical achievement, but the claims that it will lead to a solution to malaria in five years are absurd. Malaria is enormously complex and the investigators who announced their discovery seemed to have no understanding of that. Veteran malaria expert Brian Taylor pointed out that, far from a single type of mosquito being a malaria vector, there were more than 65 known species involved. The molecular biologists did not even seem to be aware of this elementary fact.

> What has been the clinical impact of this molecular knowledge? Precisely nothing

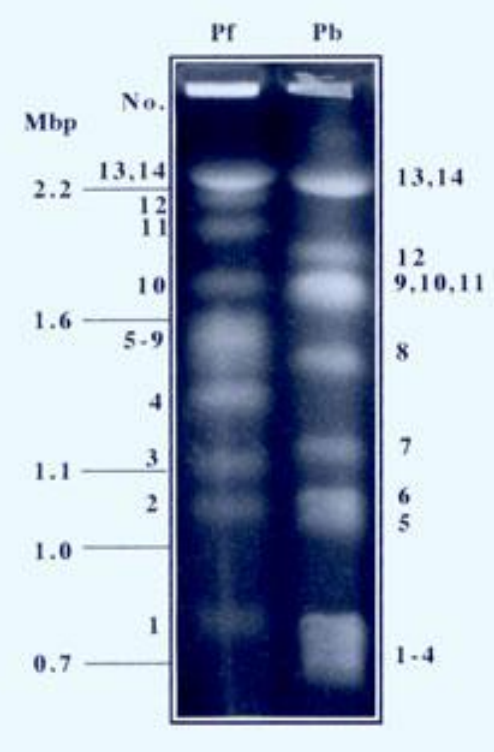
Those familiar with medical research funding know the disgraceful campaigns waged in the 70s and 80s by scientists hunting the genes for such diseases as cystic fibrosis. Give us the money, we'll find the gene and then your problems will be solved, was the message. The money was found, the genes were found — and then came nothing but a stunned contemplation of the complexity of the problem, which many clinicians had understood all along. The idea that genomics is going to make a major contribution to human health in the near future is laughable. But the tragedy is that the whole-organism biologists and clinicians who might have helped to unravel the complexity have almost all gone, destroyed by the reductionists.

If genomics is to deliver even a fraction of the promised benefits to human health, a balanced research effort must be restored, and we must drastically reduce the proportion of the available funds devoted to molecular biology.

Dr David Horrobin is author of The Madness of Adam and Eve: How Schizophrenia Shaped Humanity. This is an edited version of his essay in Frontiers 03, new writing on cutting-edge science, edited by Tim Radford. To order a copy, for £10.99 with free UK p&p, call 0870 066 7850. agreen@lavdale.co.uk

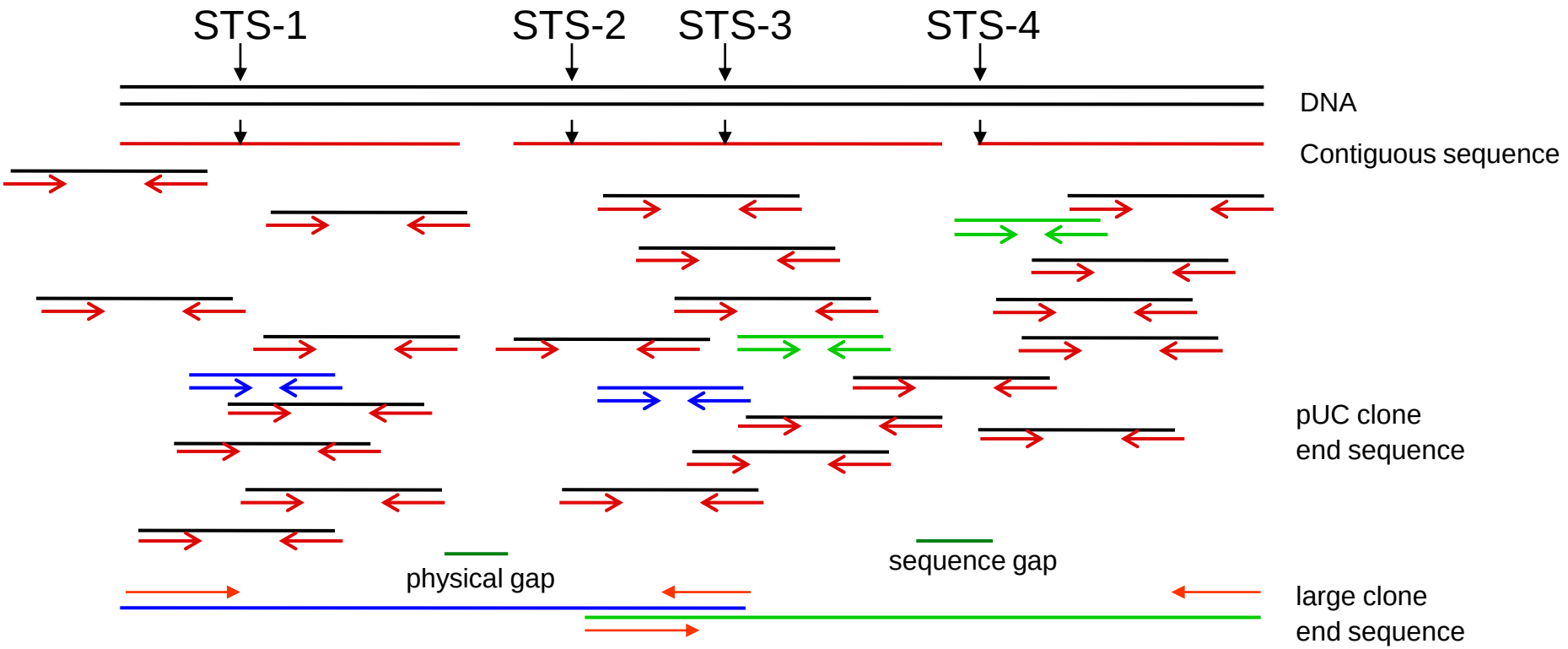
The Times (UK) 2002

Strategy



- Separate chromosomes by PFGE.
- Shotgun sequence individual chromosomes
- Align Contigs to map and close gaps using PCR/primer walking.

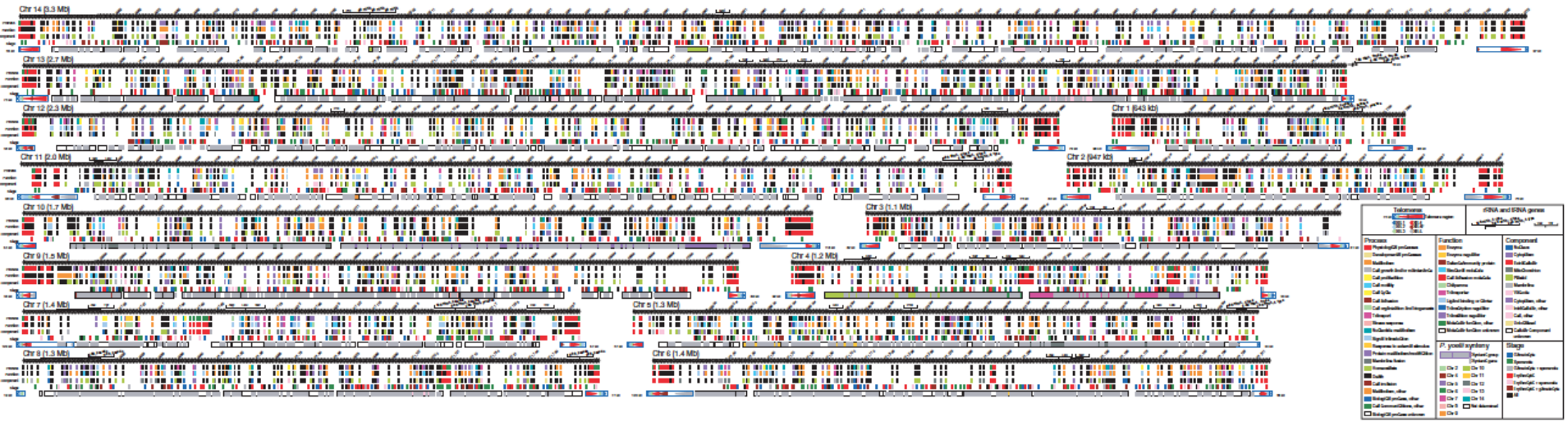
Shotgun Sequencing (back in the day)



14x coverage,
YAC clones shotgun sequenced to anchor contigs

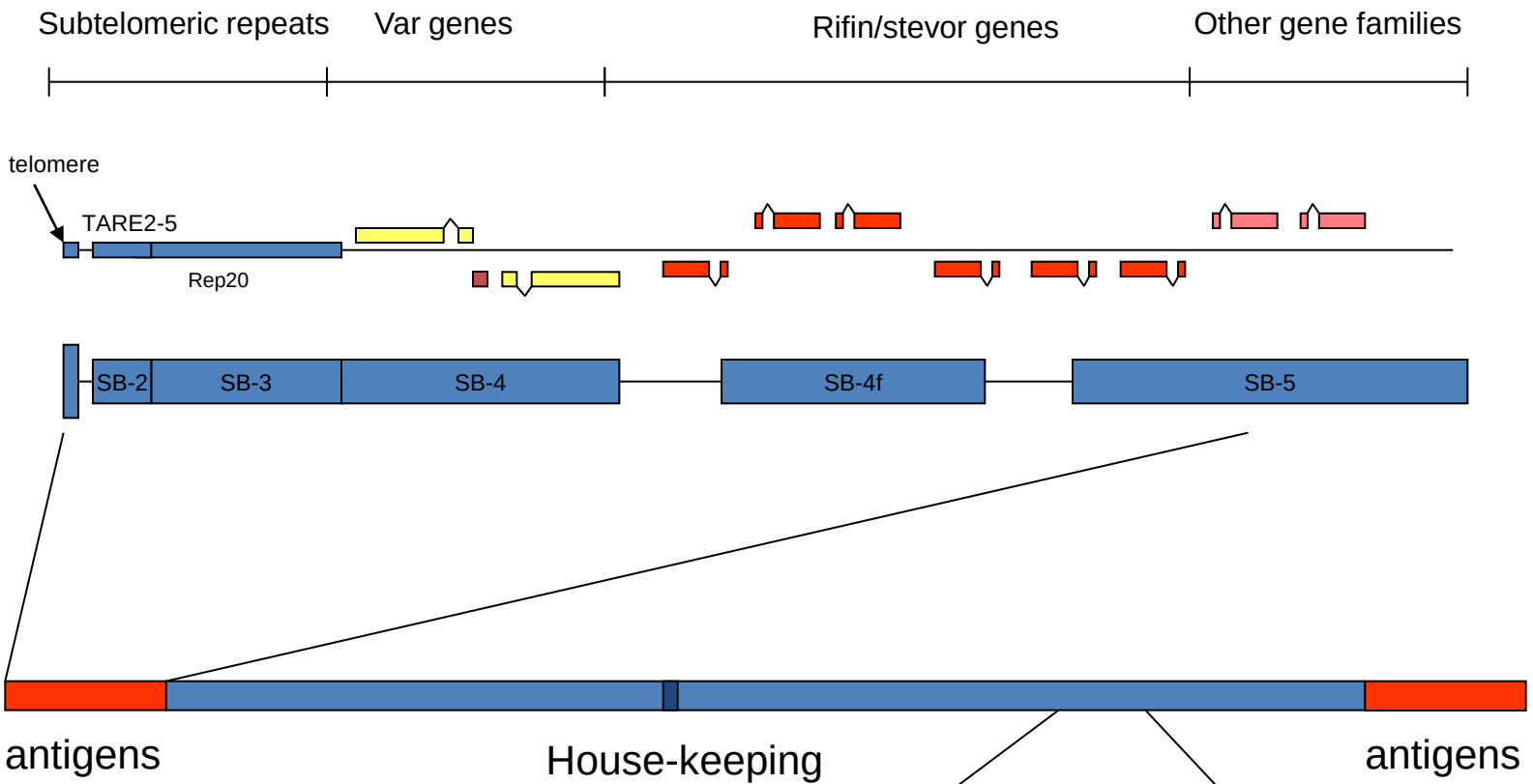
The Plasmodium falciparim genome

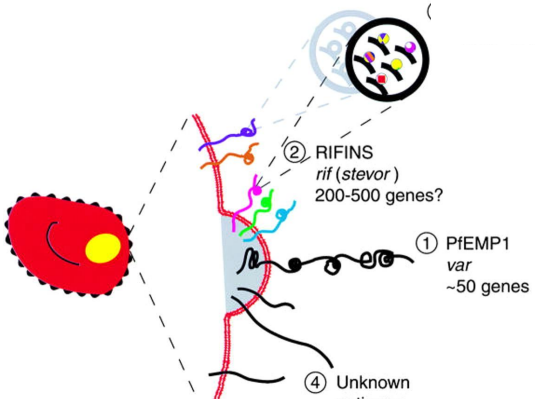
- 22853764 bp
- 19.4% GC
- 1 gene every 4.3kb (52% coding)
- Minimal set of tRNAs
- ~90% of genes have not previously described in Plasmodium.
- ~50% have no known homologues



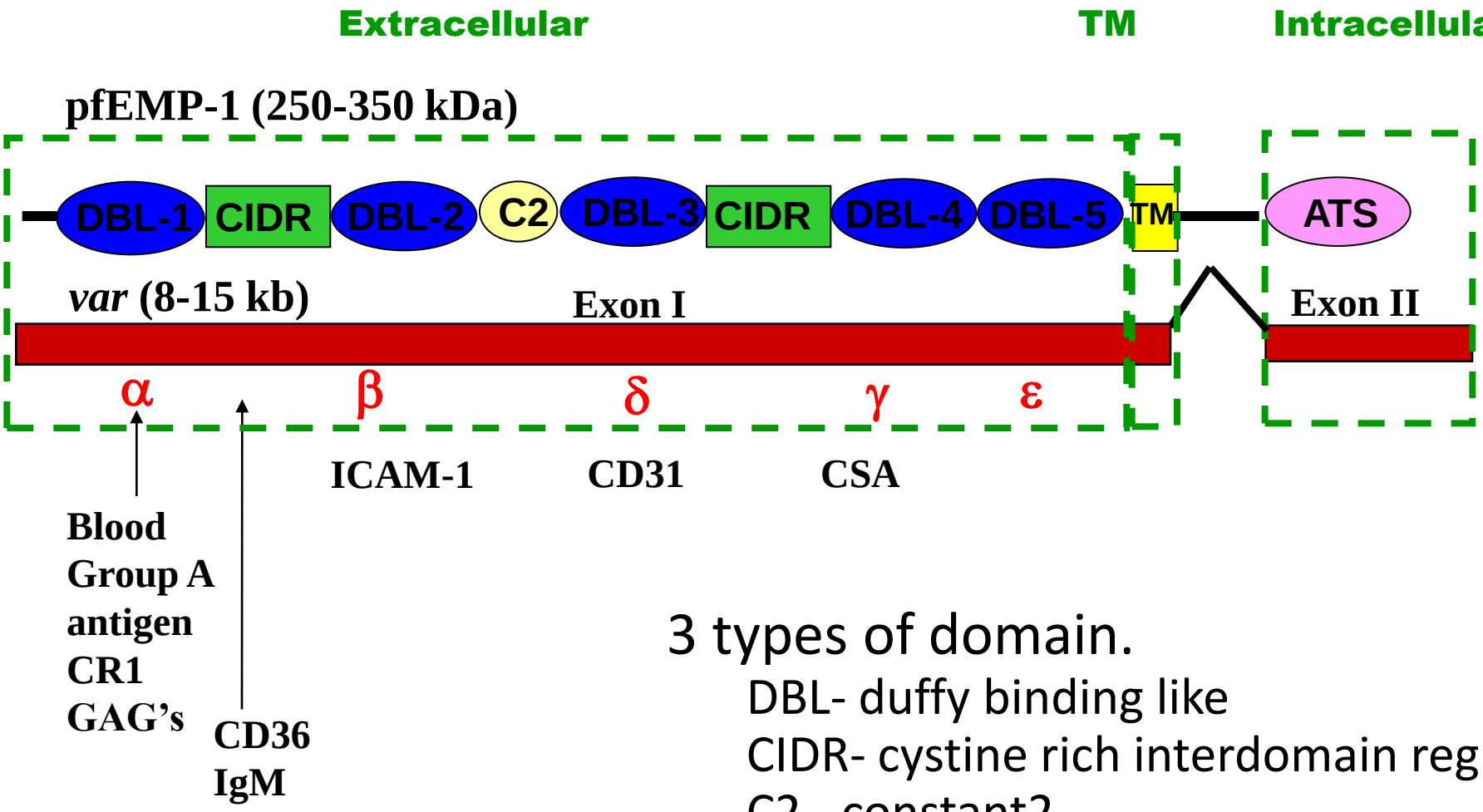
Gardener, Hall, and Fung et al (2002) Nature 419:498-511

Chromosome Structure





Adhesive domains of pfEMP-1

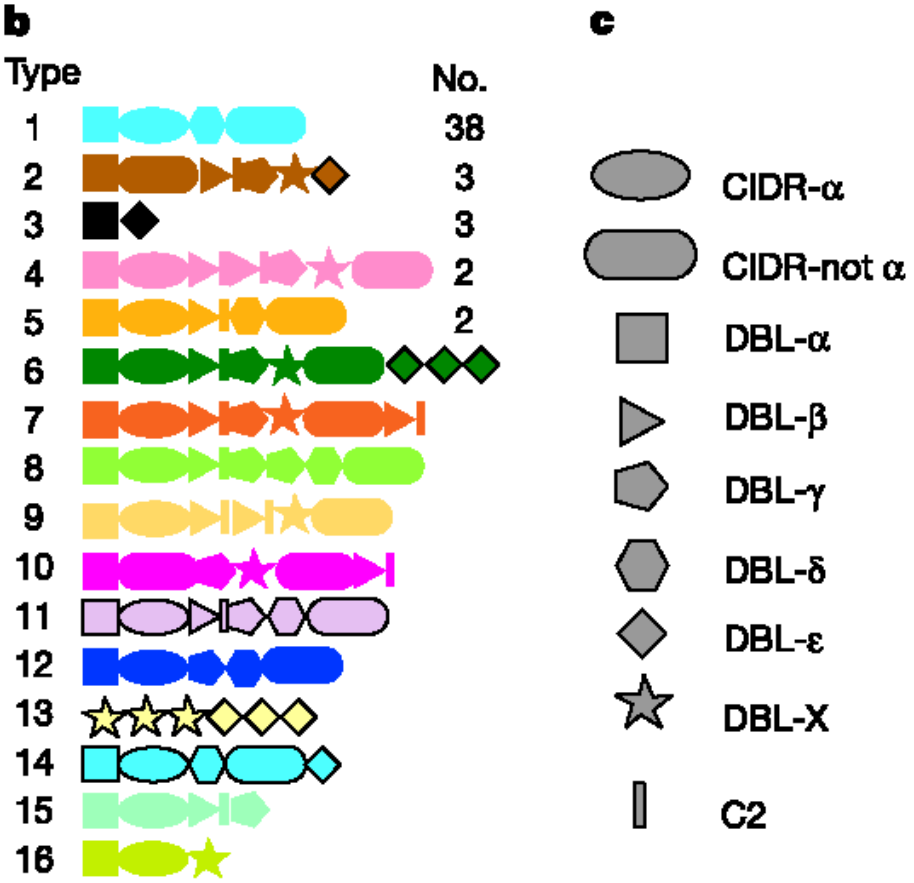


3 types of domain.

- DBL- duffy binding like
- CIDR- cystine rich interdomain regior
- C2 - constant2

VAR genes have a domain structure in Pf3D7

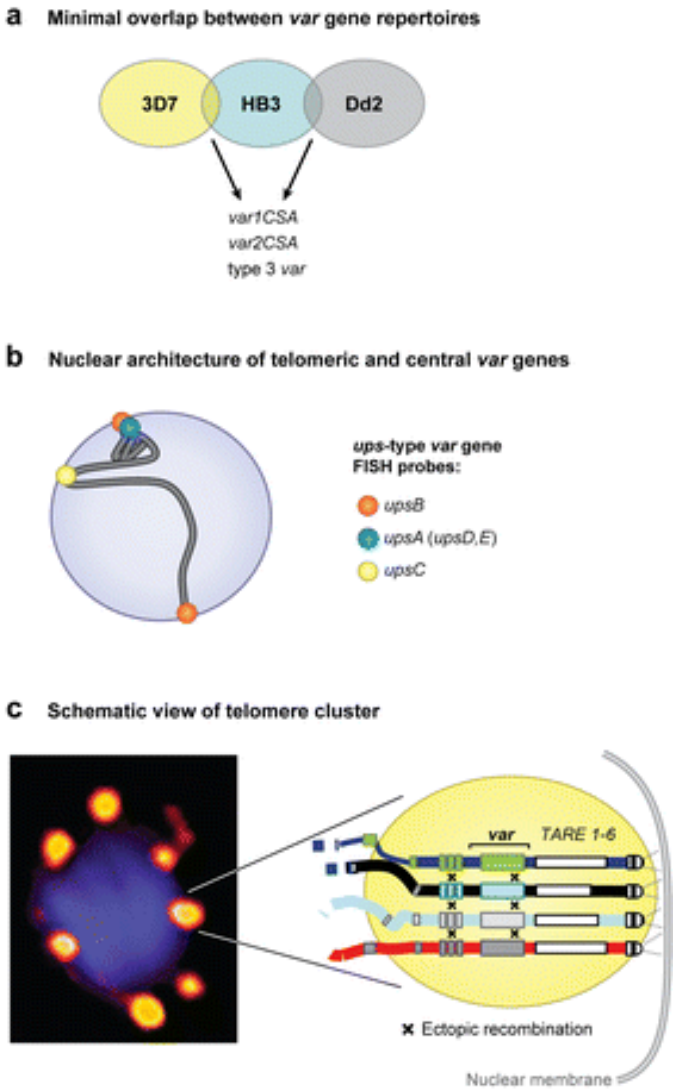
- 59 var genes in total.
- Telomeric Var genes tend to be type 1
- All internal Var clusters tend to be type 1
- Three upstream sequences
  - UPSa – 11 members- Inverted vars
  - UPSb -35 members - telomeric vars
  - UPSc 13 members - Internal vars



Gardener, Hall, and Fung et al (2002) Nature 419:498-511

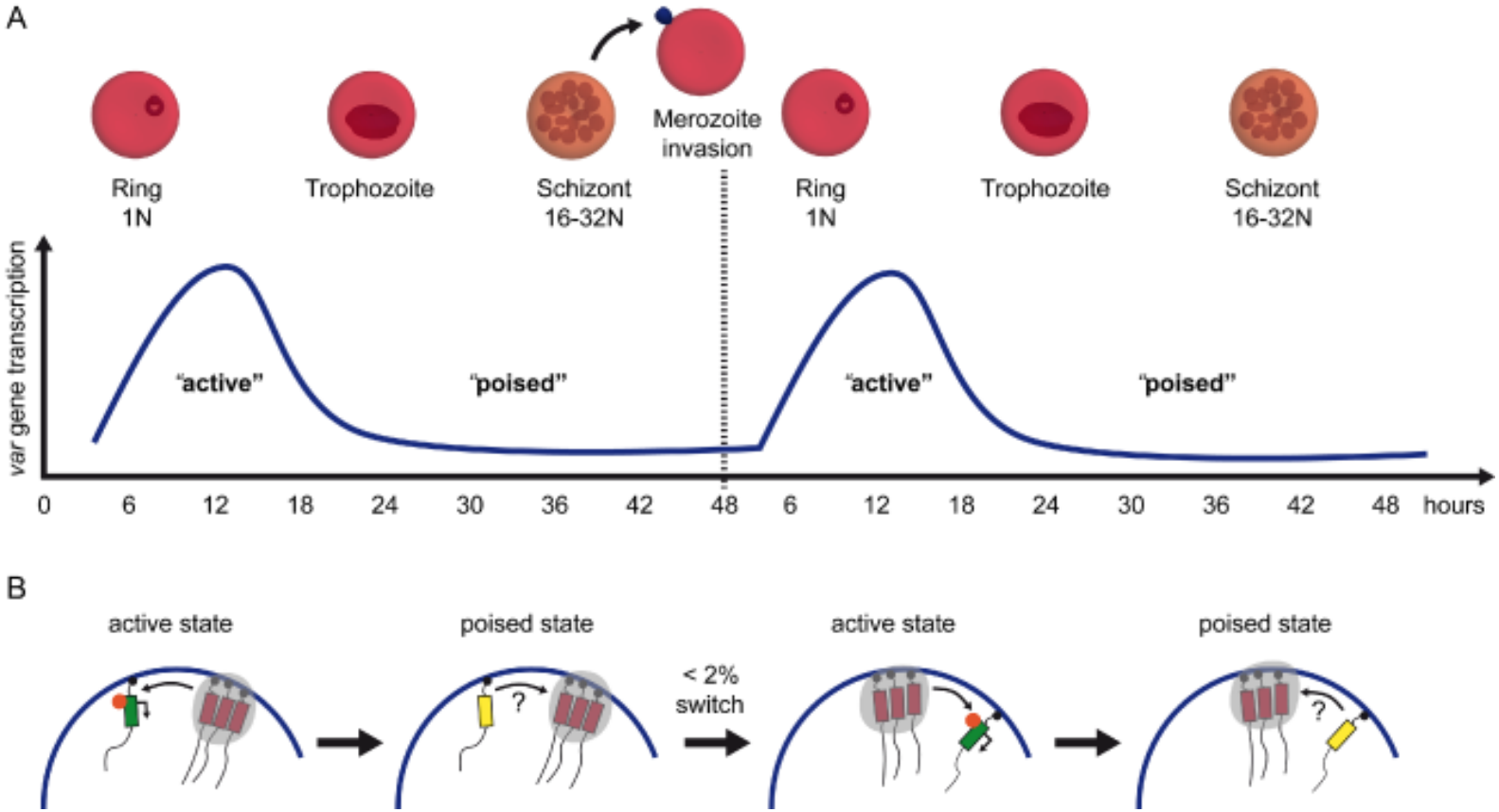
Why put these genes at telomeres?

- Telomeres are hotspots for recombination
- Telomeres co-localize in the nucleus and non-homologous recombination can occur
- Lots of gene conversion + generation of new genes.



Scherf A, et al. 2008.
Annu. Rev. Microbiol. 62:445–70

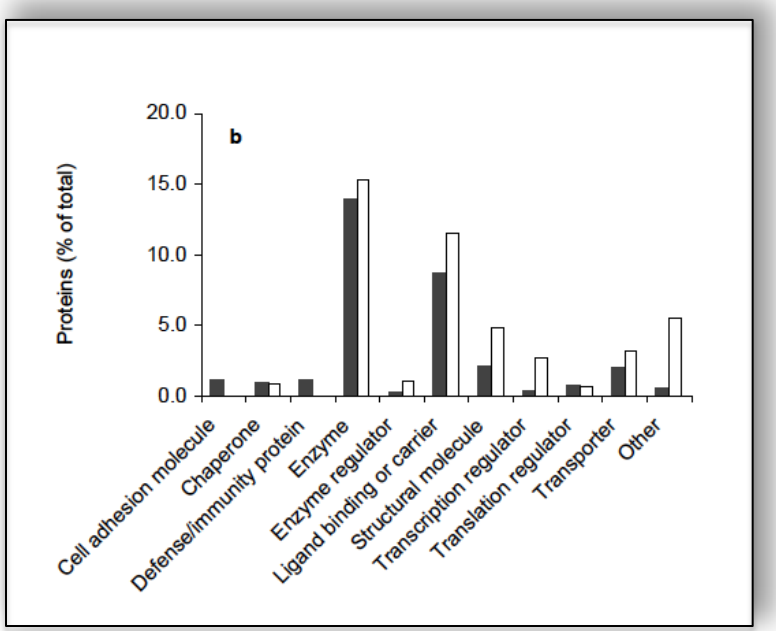
Var gene organisation is critical to antigeneic variation



UPS sequences regulate switch rate.

Guizetti J,& Scherf A. 2013 Cell Microbiol. 15(5):718-26

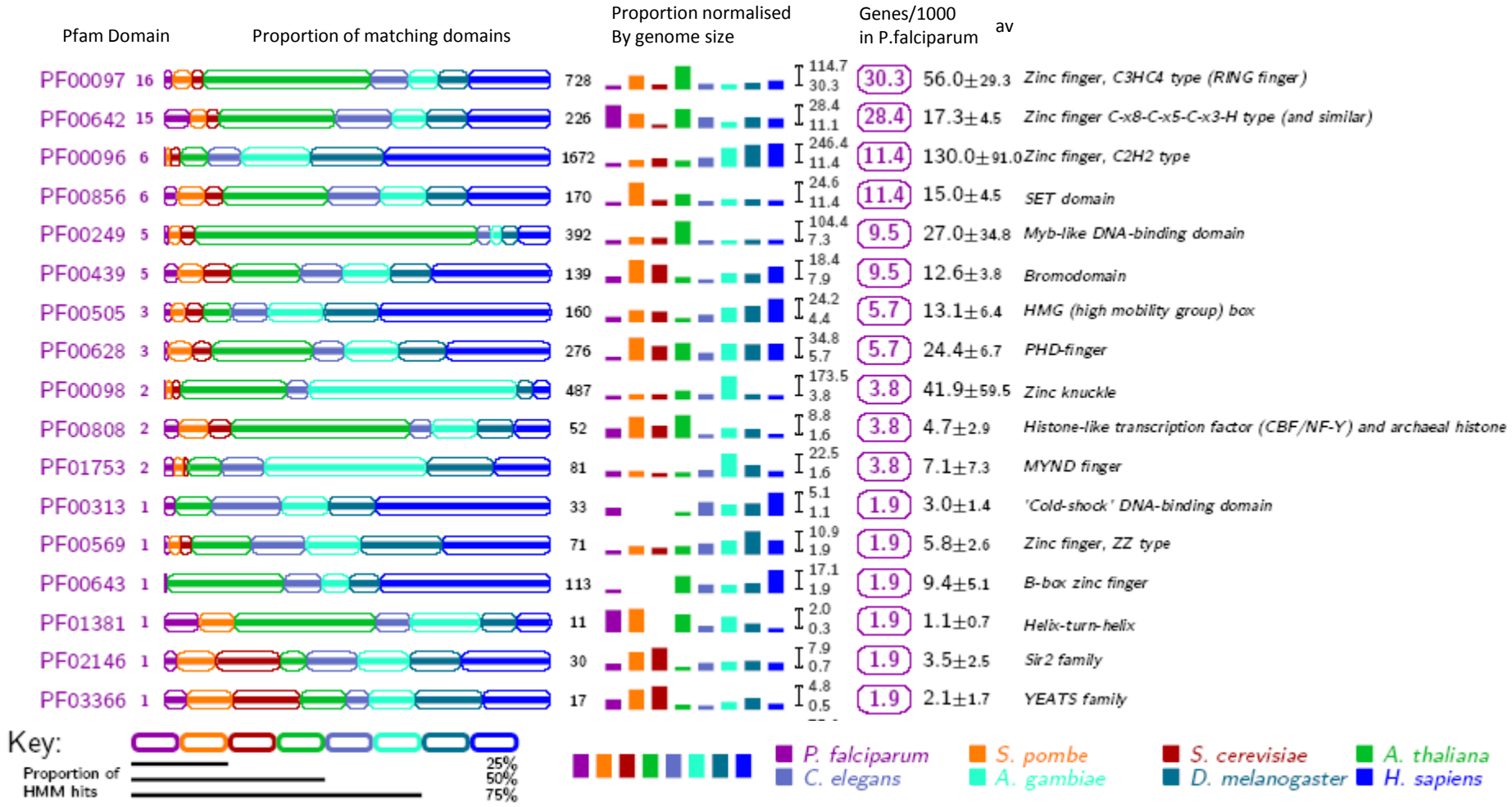
Regulation of gene expression



- *Plasmodium falciparum* has less transcription factors than you would expect in a genome of its size
  - Either they are highly diverged.
  - other mechanisms play a major roll in regulation of gene expression.

Gardner, Hall & Fung *et al* (2002) Nature 419:498-511.

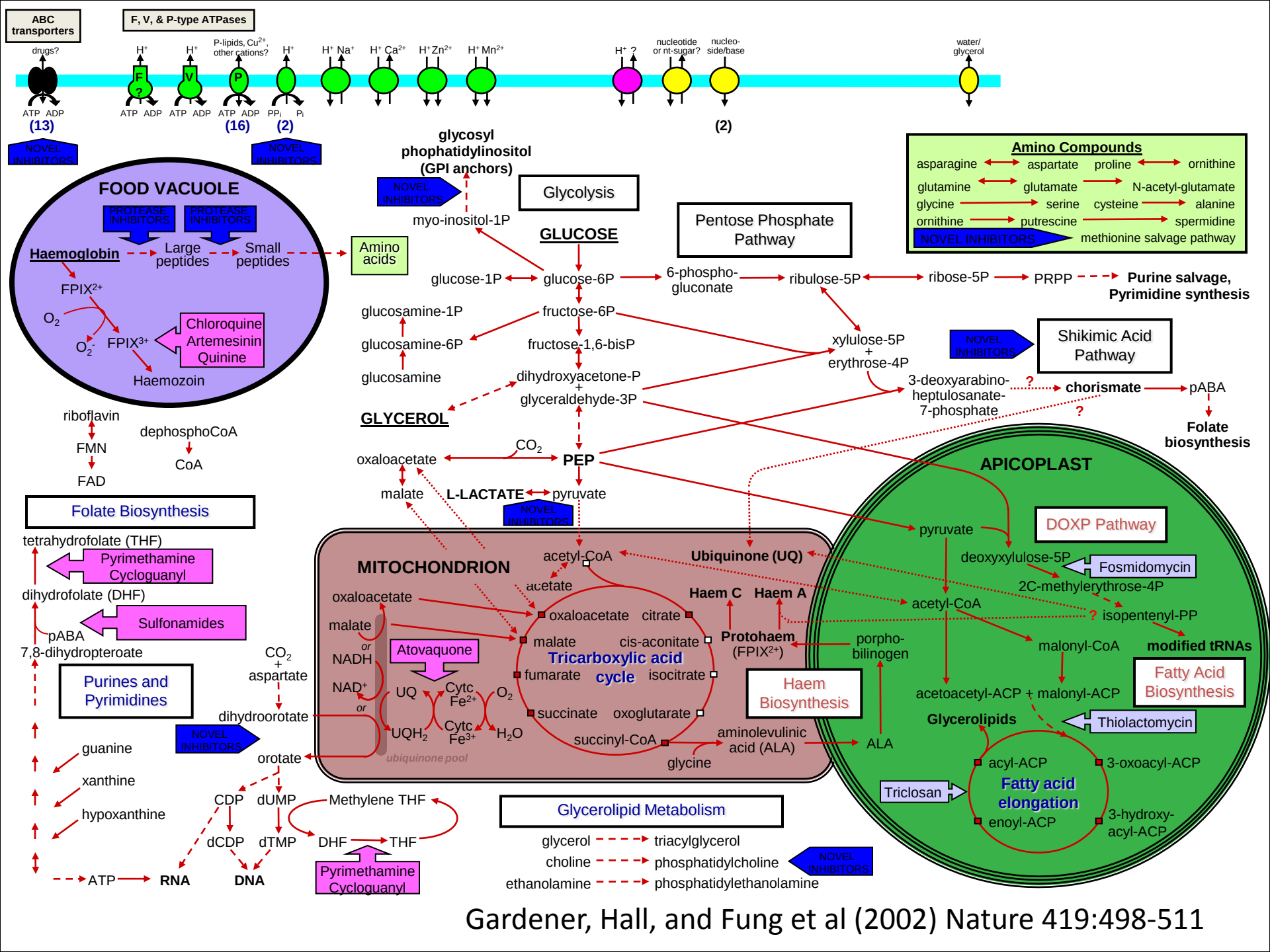
Classification of DNA binding domains in *P. falciparum*



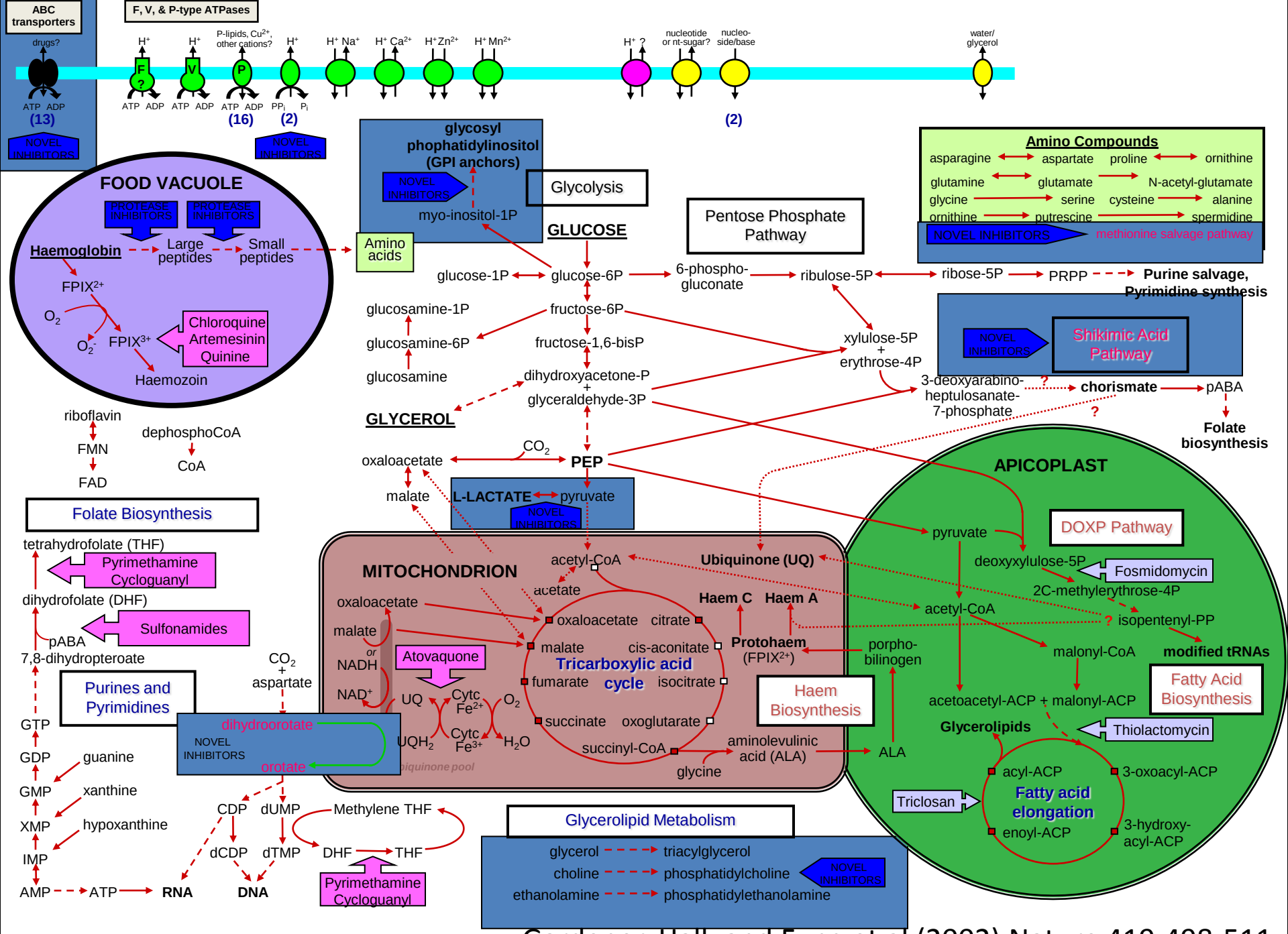
Coulson, **Hall,** Ouzounis (2004) *Genome Res. 14, 1548*

Why so few regulators?

- *Plasmodium* has a predetermined life cycle.
- It does not have nutritional “choices” as it lives in defined, stable and nutrient rich environment.
- It is not competing with other species.
- The environments it lives in are varied but occur sequentially i.e. it knows what is coming next.



Gardener, Hall, and Fung et al (2002) Nature 419:498-511



Gardener, Hall, and Fung et al (2002) Nature 419:498-511

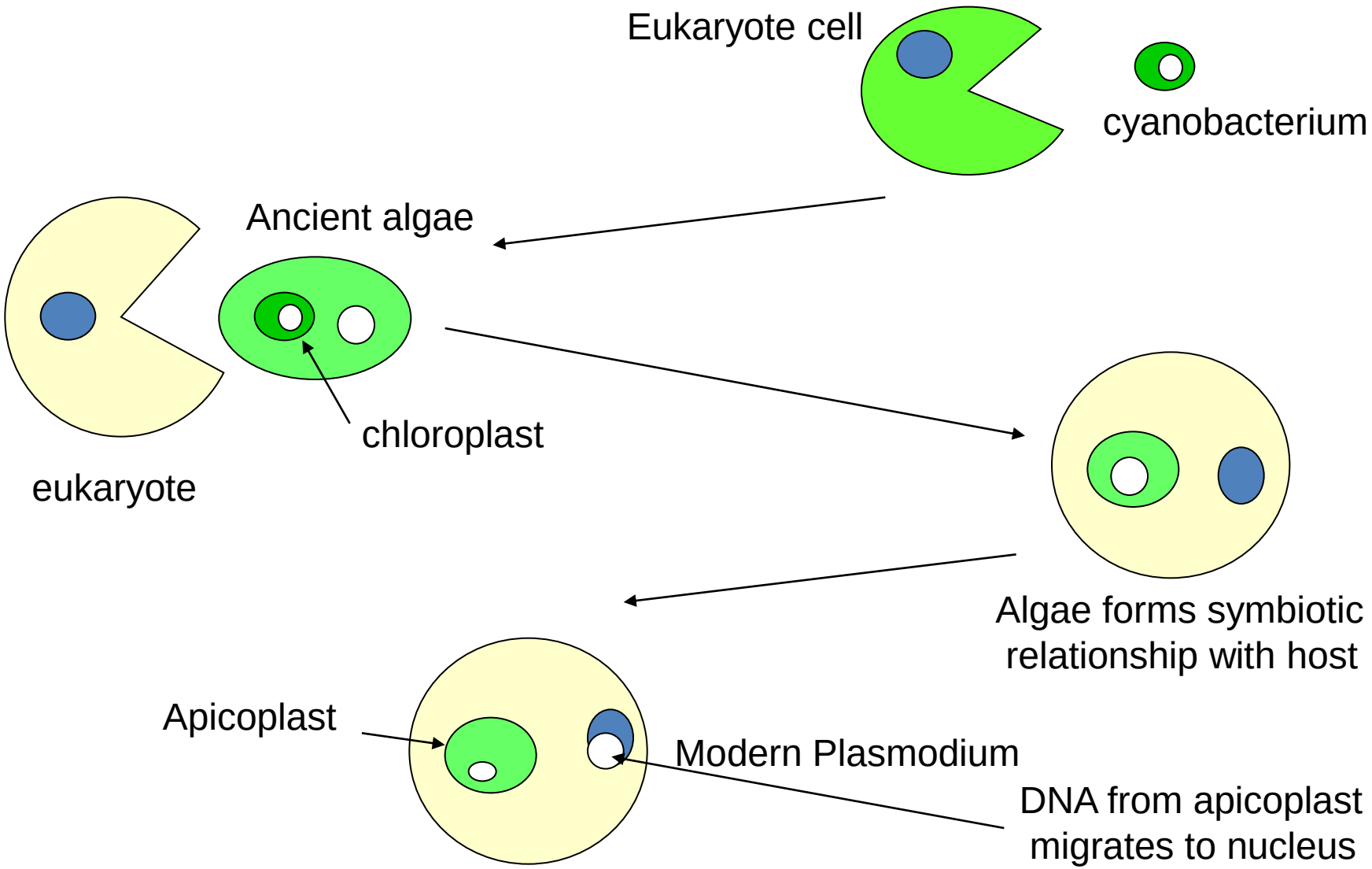
Triclosan offers protection against blood stages of malaria by inhibiting enoyl-ACP reductase of *Plasmodium falciparum*

NAMITA SUROLIA[1] & AVADHESHA SUROLIA[2]

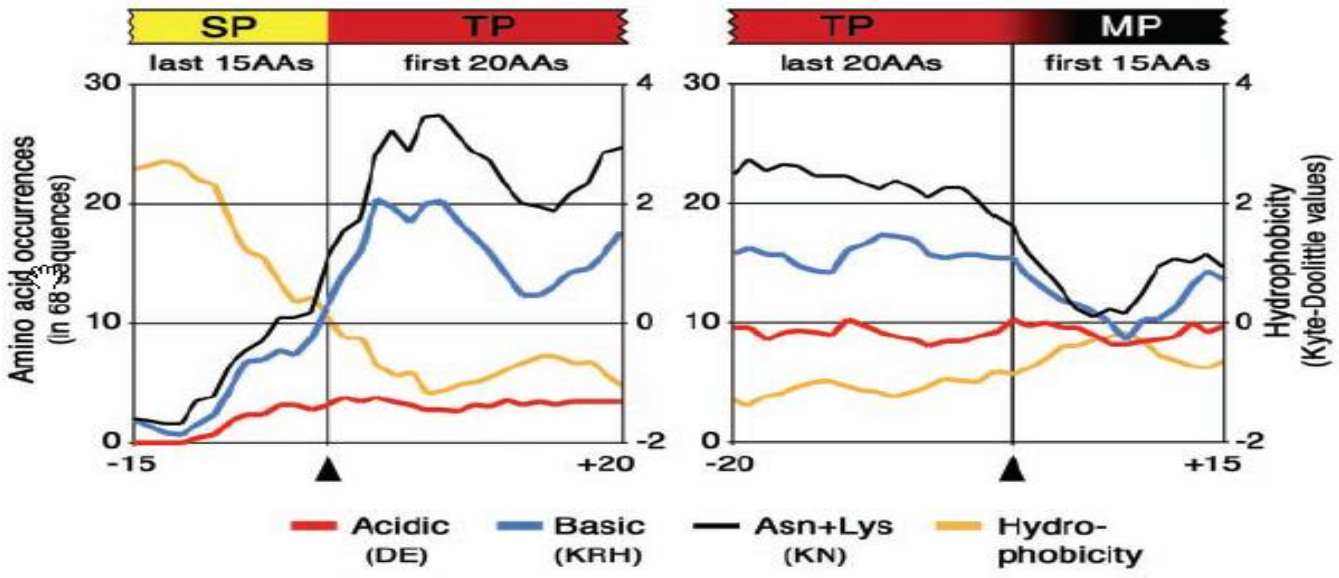
[1]*Molecular Biology and Genetics Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore, India*
[2]*Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India*
Correspondence should be addressed to N.S.; email: surolia@jncasr.ac.in

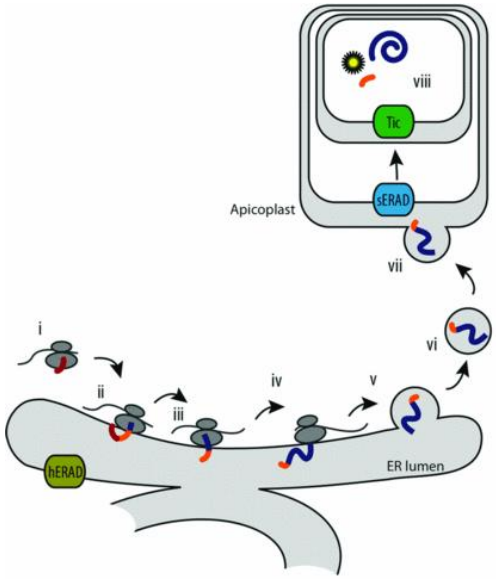
The antimicrobial biocide triclosan [5-chloro-2-(2,4-dichlorophenoxy)phenol] potently inhibits the growth of *Plasmodium falciparum in vitro* and, in a mouse model, *Plasmodium berghei in vivo*. Inhibition of [^{14}C]acetate and [^{14}C]malonyl-CoA incorporation into fatty acids *in vivo* and *in vitro*, respectively, by triclosan implicate FabI as its target. Here we demonstrate that the enoyl-ACP reductase purified from *P. falciparum* is triclosan sensitive. Also, we present the evidence for the existence of *FabI* gene in *P. falciparum*. We establish the existence of the *de novo* fatty acid biosynthetic pathway in this parasite, and identify a key enzyme of this pathway for the development of new antimalarials.

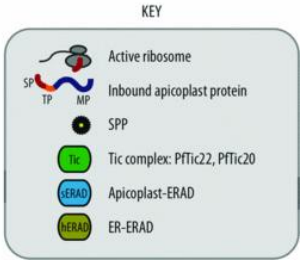
Pac Man Evolution of Malaria



Apicoplast targeting signals







~ 10% of nuclear genes may be imported into the apicoplast

Foth et al Science, 299, 705-708, 2003

Evolution

- 551 (10%) of proteins thought to be targeted to apicoplast by identification of transit peptide.
- 246 targeted to mitochondria.
- Phylogenetic analysis of 333 putative plastid targeted proteins
    - 26 plastid derived
    - 35 mitochondrial
    - 85 possibly mitochondrial

- Redirection of targeted proteins ?

- “a range of herbicides that target plastid metabolism of undesired plants are also parasiticidal, making them potential new leads for antimalarial drugs” Kalanon and McFadden (2012)

What has the genome achieved?

David Horrobin Enthusiasts for genomics have co rupted scientific endeavour and undermined hopes of medical progress

Not in the genes

Triumphalism about molecular biology, genomics and the human genome project is an increasingly pervasive theme in biomedical science. Beginning soon after Watson and Crick deciphered the structure of DNA, it has become progressively more dominant, reaching a crescendo over the past five years. Popular science magazines and the general media have united in a chorus of praise concerning the supposedly dramatic effects this is going to have on human health. But genomics enthusiasts ha e corrupted scientific endeavour and destroyed real hope of progress.

From the 1930s to the 1960s, biomedical science bore some resemblance to an integrated whole. There were researchers working at every level of biological organisation — from sub-cellular biochemistry, to whole cells, to organs, to animals, to humans. This was a golden age. Anyone receiving the best medical care in 1965 was incomparably better off than anyone in 1930.

Solutions came from many different levels. Public health specialists working at social, cultural, educational and economic levels were largely responsible for the control of malaria. Cell culture specialists working with clinicians solved polio. Microbiologists working with chemists and clinicians eliminated the threat posed by most infections. Clinicians and pharmacologists working with chemists produced the revolution in psychiatry which enabled so many patients to leave hospitals. In every field, progress depended on a constant exchange of knowledge.

But starting in the 1960s, molecular biologists and genomics specialists took over biomedical science. Everything was to be understood completely at the molecular genomic level. Everything wa to be reduced to the genome. Journals and grant-giving bodies came to be dominated by reductionists who were scathing about the complexity of whole-organ, whole-animal and especially whole-human studies which were seen as to full of uncontrolled variability to be interpretable. Clinical and physiological studies lost out and progressively their research communities were destroyed. Now we have an almost wholly reductionist biomedical community which repeatedly makes exaggerated claims about how it is going t revolutionise medical treatment — and which repeatedly fails to achieve anything.

The first genetic disease to be fully defined in molecular terms was sickle cell disease, the abnormality of the haemoglobin in human red blood cells which causes such devastation in African-origin communities. A single abnormality in a single protein causes the trouble. The abnormal protein was identified in the 1940s, the precise molecular defect was identified in the 1950s, and the three-dimensional structure of the protein was defined in the 1960s. Yet what has been the clinical impact of this wonderfully precise molecular knowledge? Precisely nothing. The clinical picture of the disease cannot yet be understood in terms of the molecular biology.

In only two areas — stomach and duodenal ulceration and organ transplantation – has medical treatment improved dramatically. In almost every other field, people are little better off than those receiving the best treatment in 1965. Take cancer. We have made dramatic progress in a very limited range of rare cancers: the leukaemias, lymphomas and testicular tumours. But even there many of the drugs used were introduced before 1965. We have simply learned to use them better.

Another example is the hype surrounding the malaria genome project. This is indeed a brilliant technical achievement, but the claims that it will lead to a solution to malaria in five years are absurd. Malaria is enormously complex and the investigators who announced their discovery seemed to have no understanding of that. Veteran malaria expert Brian Taylor pointed out that, far from a single type of mosquito being a malaria vector, there were more than 65 known species involved. The molecular biologists did not even seem to be aware of this elementary fact.

> What has been the clinical impact of this molecular knowledge? Precisely nothing

Those familiar with medical research funding know the disgraceful campaigns waged in the 70s and 80s by scientists hunting the genes for such diseases as cystic fibrosis. Give us the money, we'll find the gene and then your problems will be solved, was the message. The money was found, the genes were found — and then came nothing but a stunned contemplation of the complexity of the problem, which many clinicians had understood all along. The idea that genomics is going to make a major contribution to human health in the near future is laughable. But the tragedy is that the whole-organism biologists and clinicians who might have helped to unravel the complexity have almost all gone, destroyed by the reductionists.

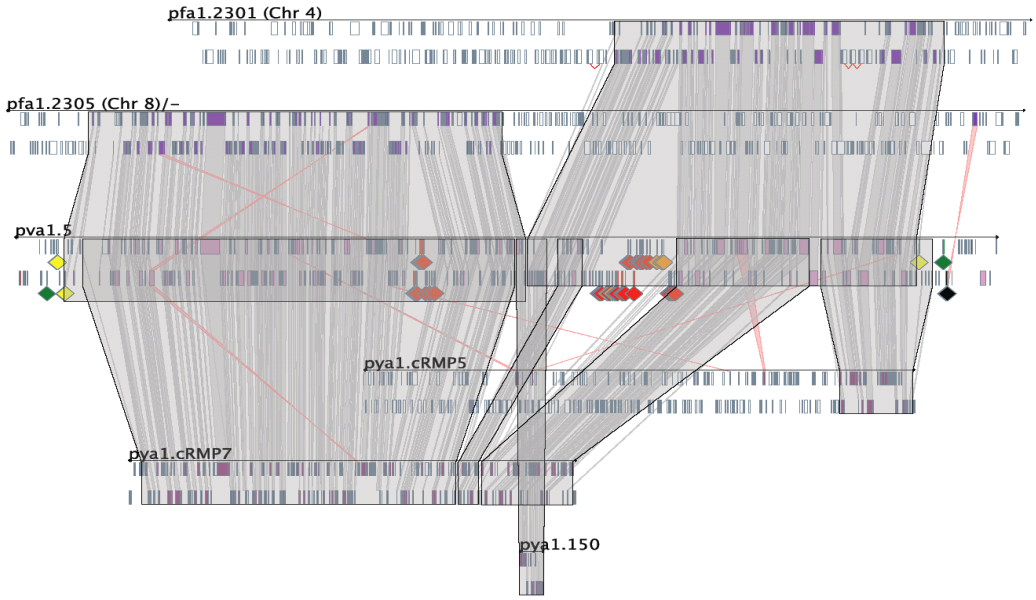
If genomics is to deliver even a fraction of the promised benefits to human health, a balanced research effort must be restored, and we must drastically reduce the proportion of the available funds devoted to molecular biology.

Dr David Horrobin is author of The Madness of Adam and Eve: How Schizophrenia Shaped Humanity. This is an edited version of his essay in Frontiers 03, new writing on cutting-edge science, edited by Tim Radford. To order a copy, for £10.99 with free UK p&p, call 0870 066 7850.

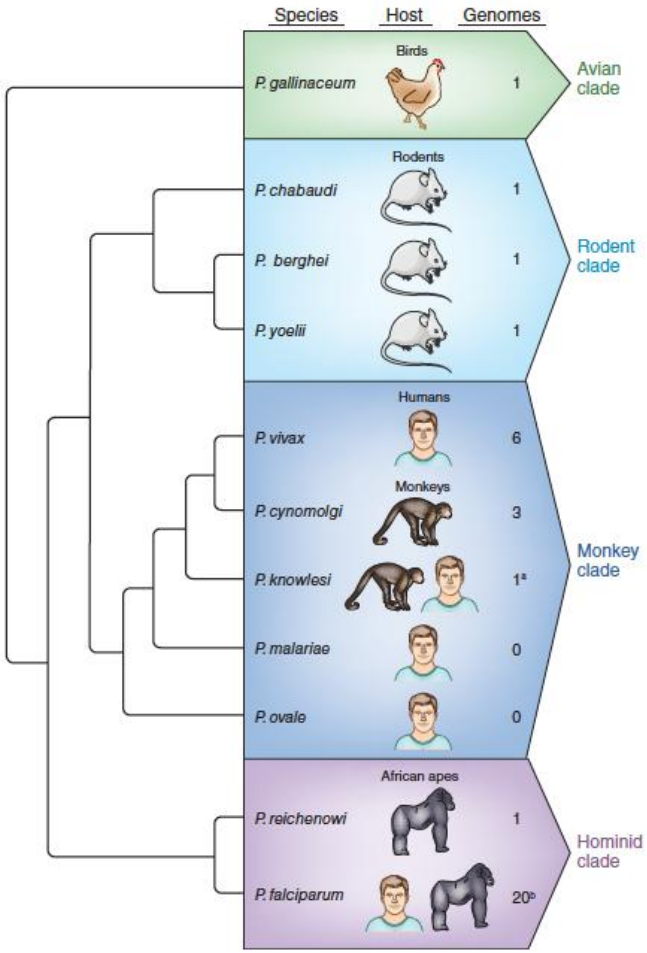
agreen@laxdale.co.uk

- A detailed understanding of the molecular basis antigenic variation
- A metabolic map for drug target identification
- A list of genes targeted to the apicoplast a known drug target.

Comparative Genomics of Malaria Parasites



Plasmodium species are highly host restricted

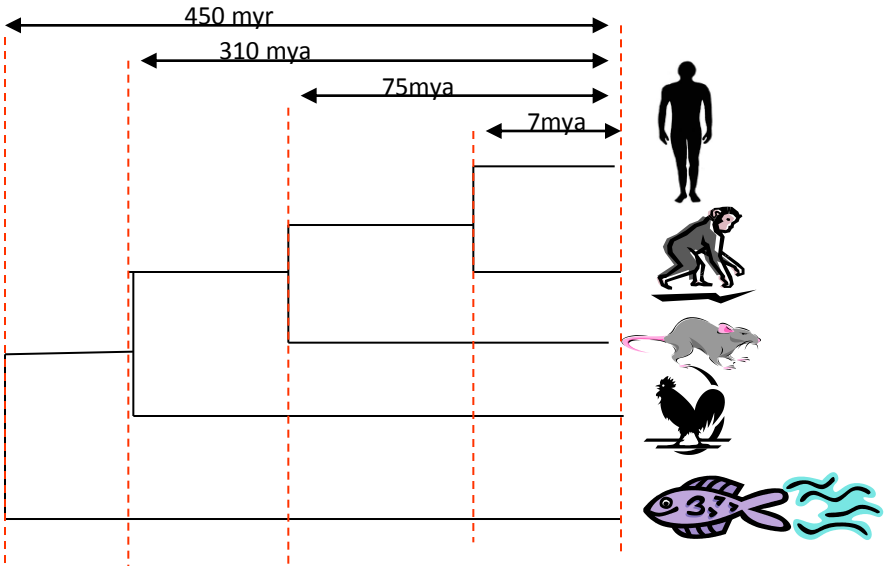


- Host jumping is rare
  - *P. Knowlesi* into human
- Most plasmodium parasites are highly adapted to a single host or a few closely related species.

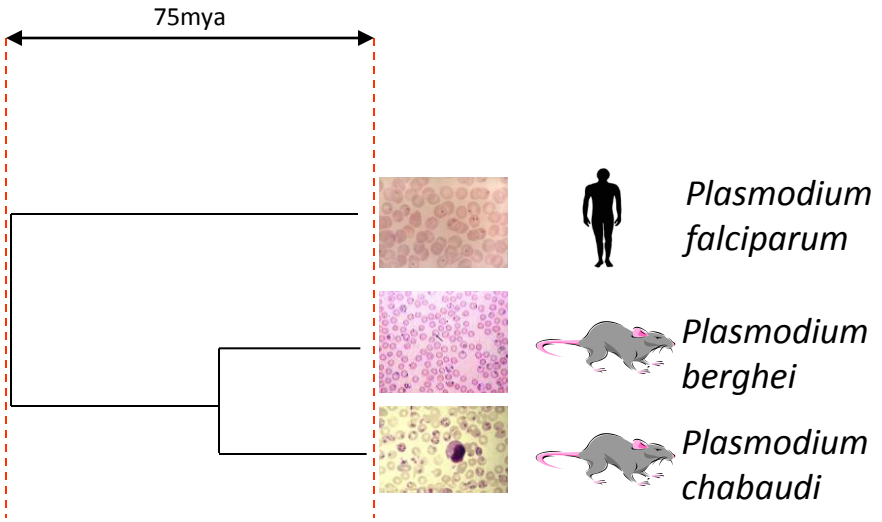
Hall, N Nature genetics (2013)

Eukaryotic Evolution

Chordate evolution



Plasmodium evolution

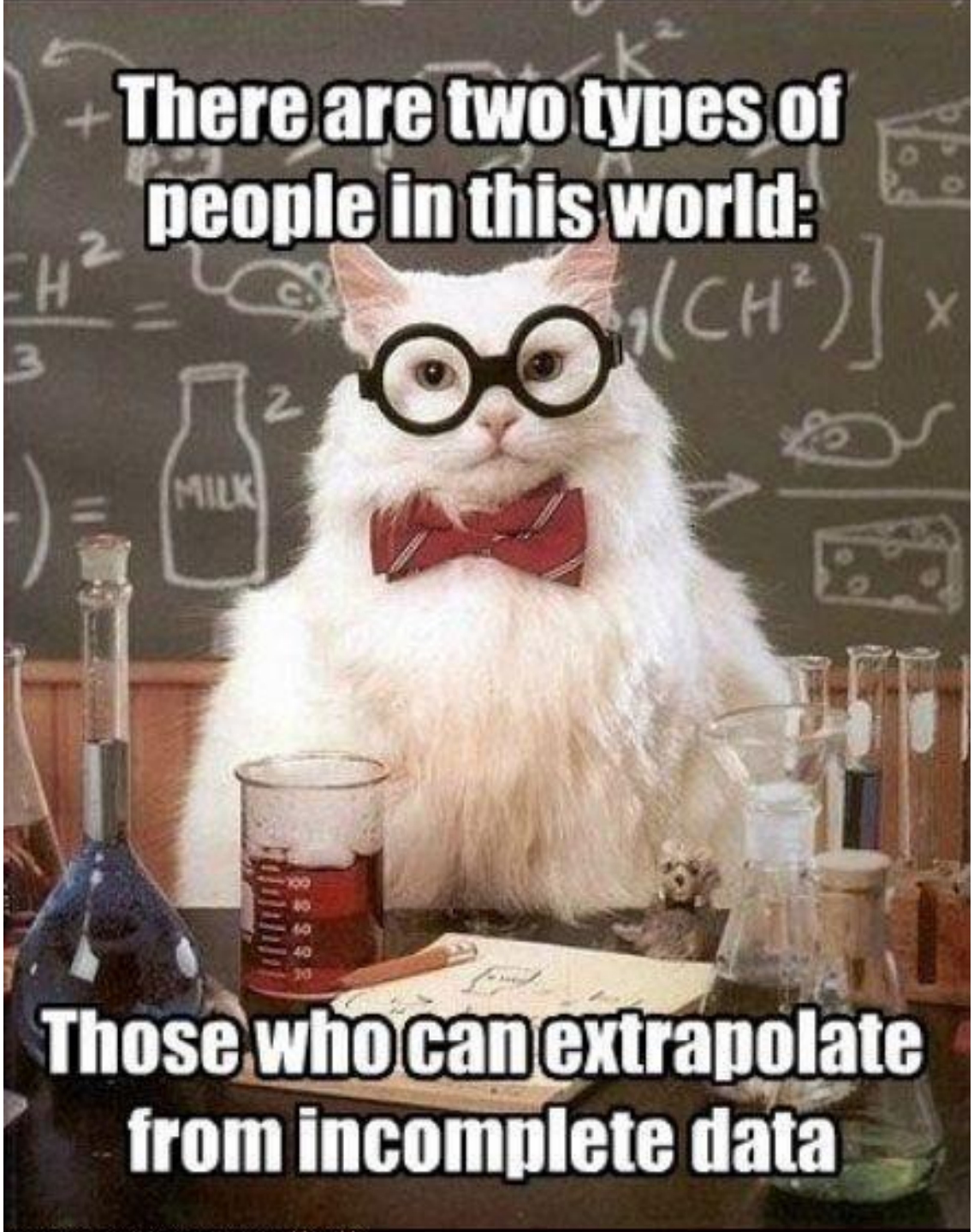


Overview of assembly and gene Count

	P. berghei	*P. c. chabaudi*	*P. y. yoelii* *	*P. falciparum*
Size (bp)	17,996,878	16,866,661	23,125,449	22,853,764
No. contigs	7497	10679	5687	93
Av. contig size (bp)	2,400	1,580	4,066	213,586
Sequence coverage	4x	4x	5x	14.5x
No. protein coding genes	5864	5698	5878	5268

Hall et al (2005) Science 307:82-6

*Carlton (2002) Nature 419:512-9.


There are two types of people in this world:
MILK
Those who can extrapolate from incomplete data

What is a good N50?….it depends

ARTICLE

doi:10.1038/nature11650

Analysis of the bread wheat genome using whole-genome shotgun sequencing

Rachel Brenchley[1], Manuel Spannagl[2], Matthias Pfeifer[2], Gary L. A. Barker[3], Rosalinda D'Amore[1], Alexandra M. Allen[3], Neil McKenzie[4], Melissa Kramer[5], Arnaud Kerhornou[6], Dan Bolser[6], Suzanne Kay[1], Darren Waite[4], Martin Trick[4], Ian Bancroft[4], Yong Gu[7], Naxin Huo[7], Ming-Cheng Luo[8], Sunish Sehgal[9], Bikram Gill[9], Sharyar Kianian[10], Olin Anderson[7], Paul Kersey[6], Jan Dvorak[8], W. Richard McCombie[5], Anthony Hall[1], Klaus F. X. Mayer[2], Keith J. Edwards[3], Michael W. Bevan[4] & Neil Hall[1]

Table 2 | Assembly statistics of the orthologous group assembly, the LCG and cDNA assemblies

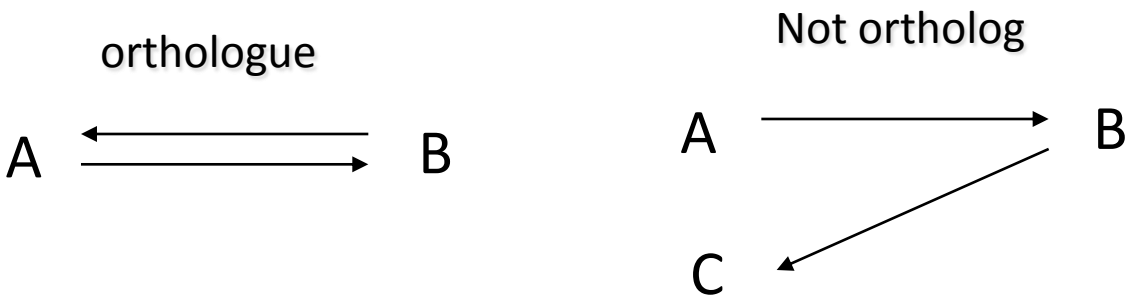
	Orthologous group assembly* (99% m.i.)	LCG†
Number of sequences	949,279	5,321,847
Total sequence (bp)	437,512,281	3,800,325,216
Minimum length; maximum length (bp)	79; 7,312	100; 21,721
N10; N50; N90 (bp)	766; 481; 331	2,234; 884; 420
Mean length (bp)	460.89	714.10
GC content (%)	48.25	47.69

Orthology, Homology, Parology

- Homology
  - Genes that are similar due to related ancestry.
- Orthologs
  - Homologous genes separated due to a speciation event
- Paralogs
  - Homologous genes separated by a duplication event

Orthologue Identification in between *Plasmodium* species

- Reciprocal orthologues identified using a BLAST cutoff of score of 50.

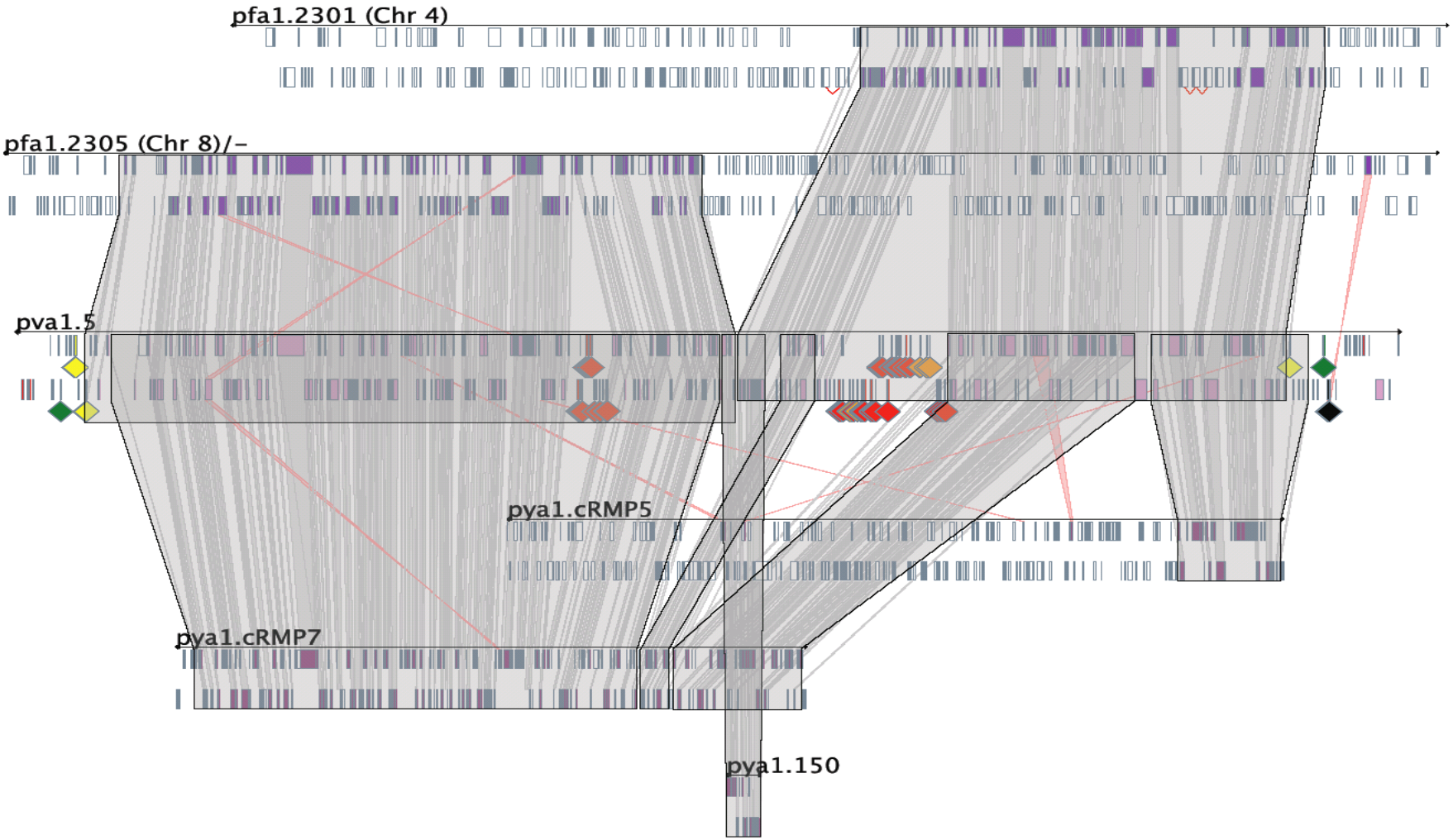


- 5268 Genes in *P. falciparum*
- *4391* have orthologs in other species
- Another 109 orthologs can be identified using synteny
- 736 have no orthologs.

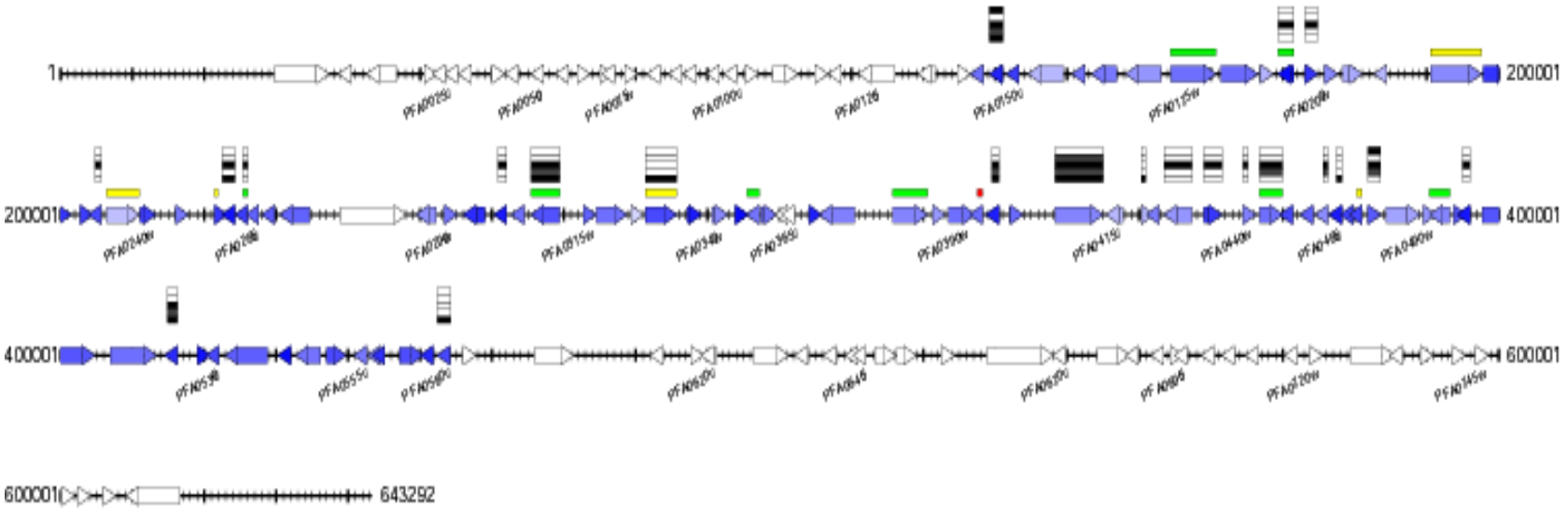
- *genes in RMP genomes clustered using Tribe-MCl*

where are the differences ?

Plasmodium genomes are syntenic



Orthologue map of *P. falciparum* Chr1



Of the 736 genes without orthologs. Only 161 are in internal regions.
Of these 45 are members of Pf specific gene families
19 are Pf specific expansions
Only 12 have predicted functions

Most (575) Species Specific genes are at the sub-telomeres

Hall et al (2005) Science 307:82-

Telomeric gene families

	P. falciparum	*P. chabaudi*	*P. berghei*	*P. yoelii*
VAR	59	-	-	-
Rifin	177	-	-	-
PIR	-	138	180	838
Pyst-a	1	108	45	168
Pyst-b	-	10	34	57
Pcst-b	-	75	11	7

Pf specific families

Novel gene families

Families clustered using Tribe-MCL

Hall et al (2005) Science 307:82-6

Evolution by gene family expansion

- There are very few genes that are truly “unique” to a *Plasmodium* species.
- There is a lot of gene family expansion that is species specific.
- This mostly occurs at telomeres.
- Some genes which are single copy in one species are multi-copy in another…and presumably have acquiring new functions.

Molecular se tion analysis

Silent and Replacement Substitutions

- ‘Housekeeping Genes’.
  - Negative (Purifying) selection.
    - Change is Bad.
- Genes that have a role in adaption.
  - Positive (Adaptive) selection.
    - Change is Good.
- Selectively neutral genes
  - Genetic drift.
    - Change does not matter

Silent substitution:

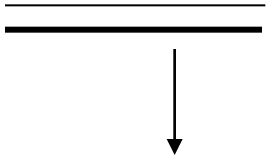
Sequence 1:	**UUU CAU CGU**
Sequence 2:	**UUU CAC CGU**
Coded Amino Acids:	***Phe His Arg***

Replacement substitution:

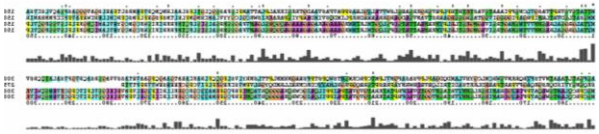
Sequence 1:	**UUU CAU CGU**
Sequence 2:	**UUU CAG CGU**
Coded Amino Acids:	***Phe His Arg*** / ***Gln***

dN/dS analysis

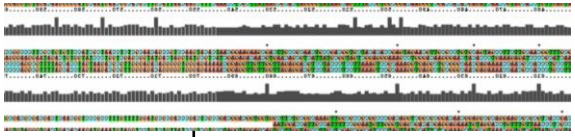
Orthologous gene



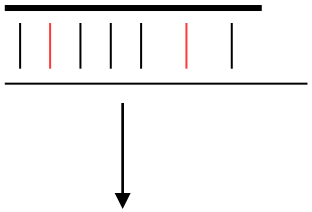
Muscle Protein alignment



nucleotide alignment



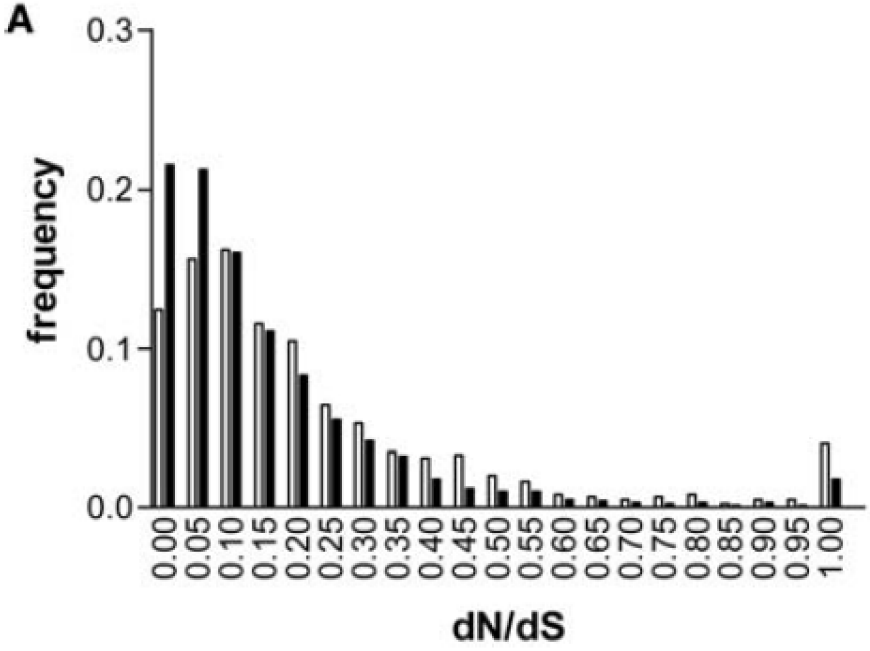
Paml calcluate model of Evolution using maximum likelihood



dN/dS evolution rate

Positive selection on plasmodium genes is stronger in the mosquito vector

A



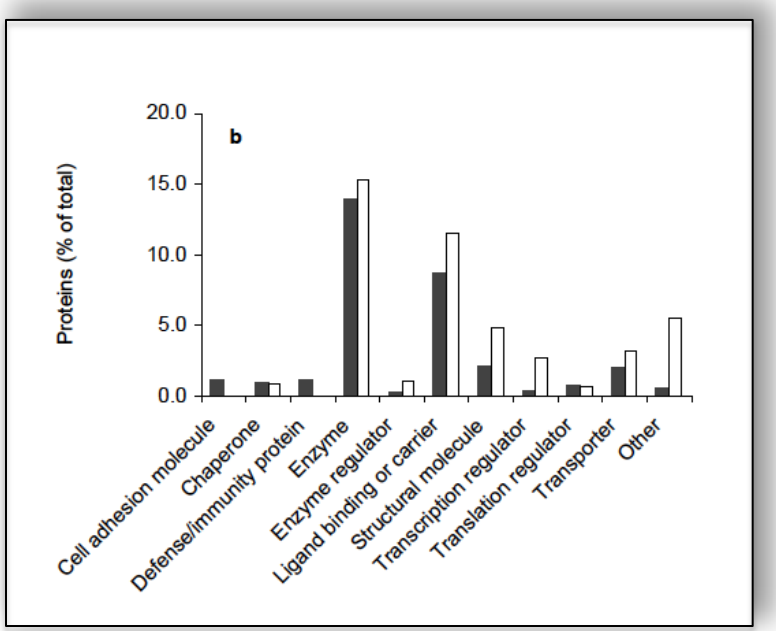
B

Orthologs	SP or TM		Non-SP or TM		D	P
	n	median	n	median		
Genomics - all stages	726	0.174	3952	0.121	0.147	>0.001*
Proteome and transcriptome - all stages	147	0.157	501	0.100	0.217	>0.001*
- mammal expressed	106	0.159	311	0.098	0.237	>0.001*
- mosquito expressed	41	0.139	188	0.104	0.17	<0.05

-The dN/dS of secretory proteins expressed in the vertebrate host is significantly higher than that of non secretory proteins

Vertebrate expressed secretory proteins have a significantly higher dn/ds compared to mosquito expressed

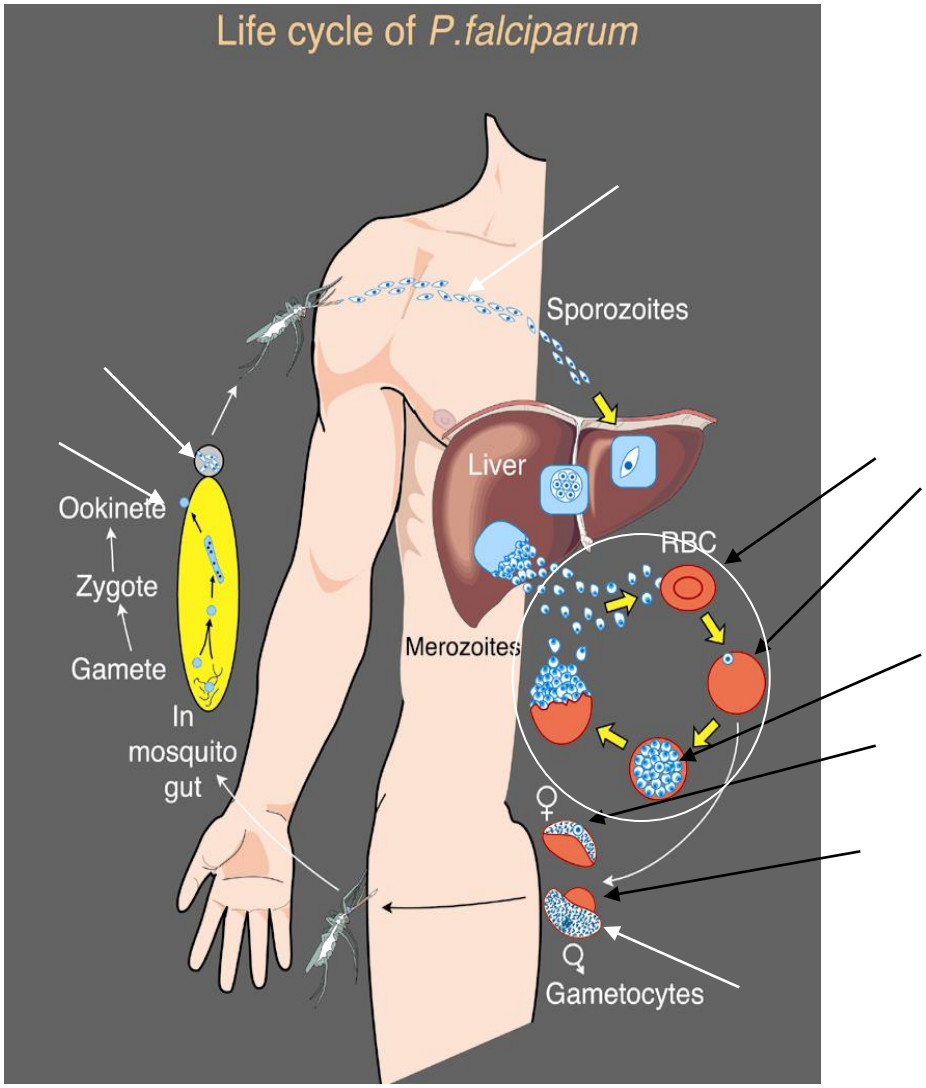
Regulation of gene expression



- *Plasmodium falciparum* has less transcription factors than you would expect in a genome of its size
  - Either they are highly diverged.
  - other mechanisms play a major roll in regulation of gene expression.

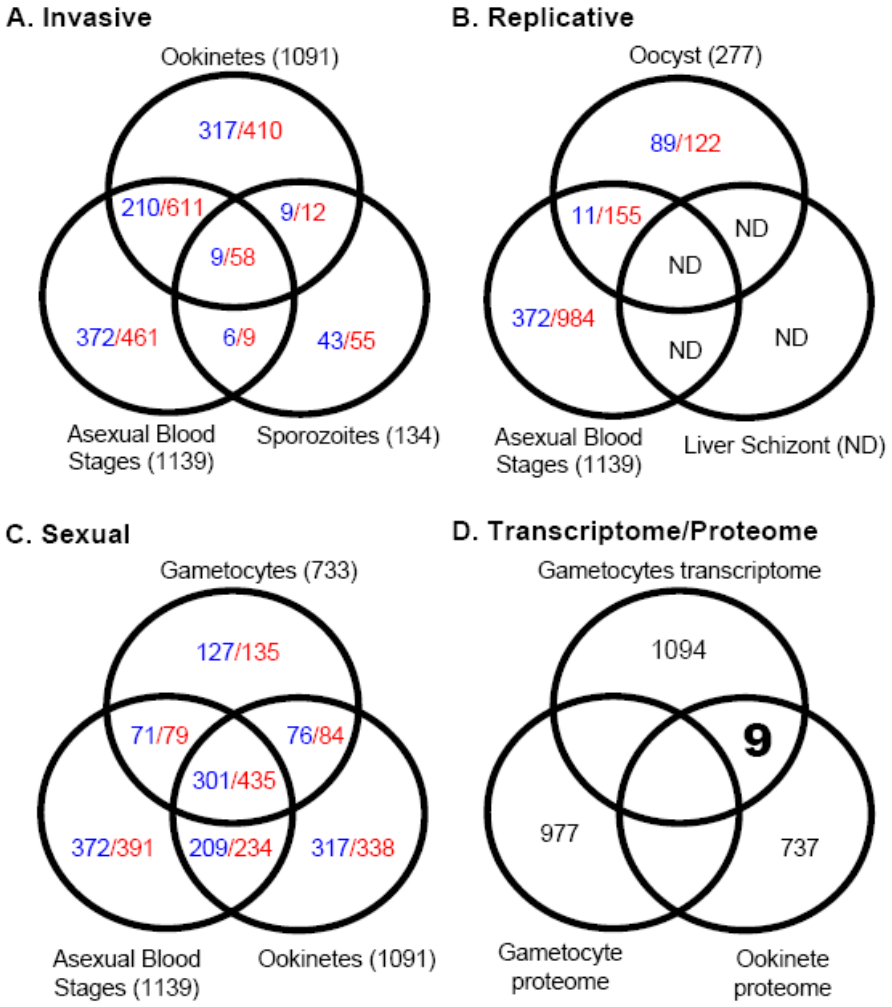
Gardner, Hall & Fung *et al* (2002) Nature 419:498-511.

Expression data for *P. berghei*



- Amplicon based microarray, data from RBC and gametocyte stages
- Proteomic data from RBC, mosquito and sporozoite stages.

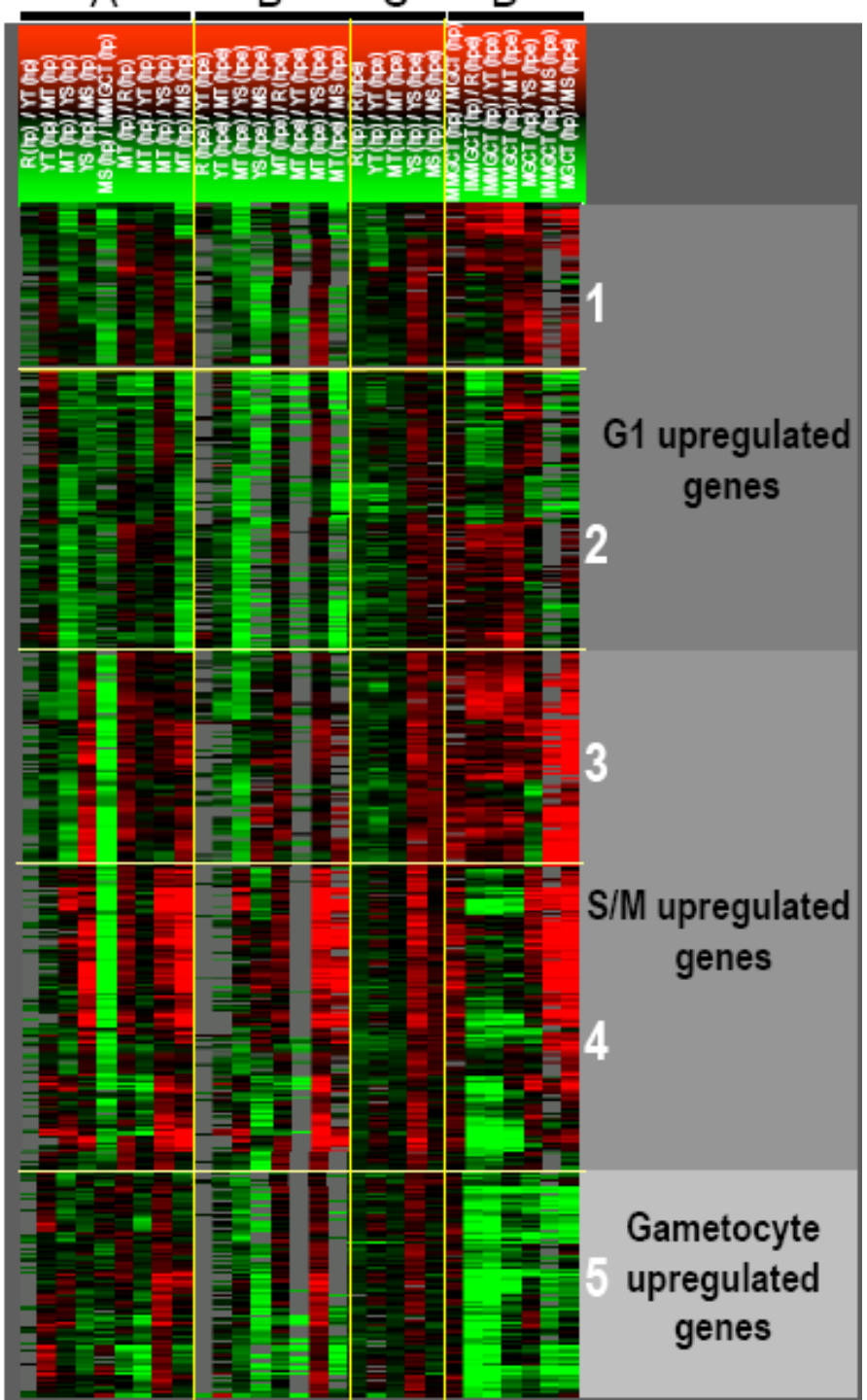
Proteomic data



Found Exclusively in this stage

Non-exclusive to stage

Hall *et al* Science (2005)



Identification of putative translational repressor motif in *P. berghei*

Identify transcripts upregulated only in the gametocyte

↓ 49

Of these find those that only have detectable protein products in ookinetes

↓ 9

Extract 3’ and 5’ UTRs and search for a conserved motif using MEME

↓

Search for Motif back against the genome

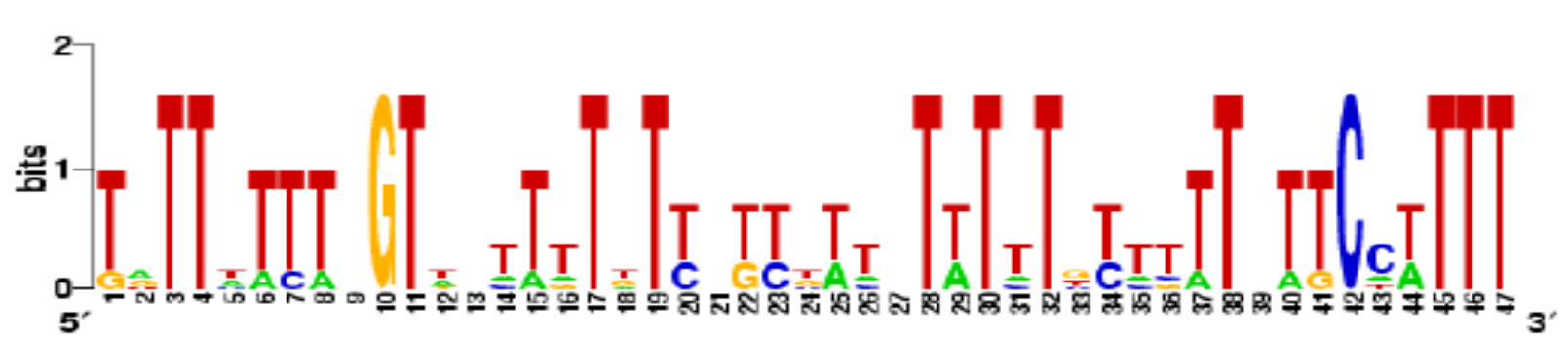
↓

Compare to expression data

Hall *et al* Science, 2005

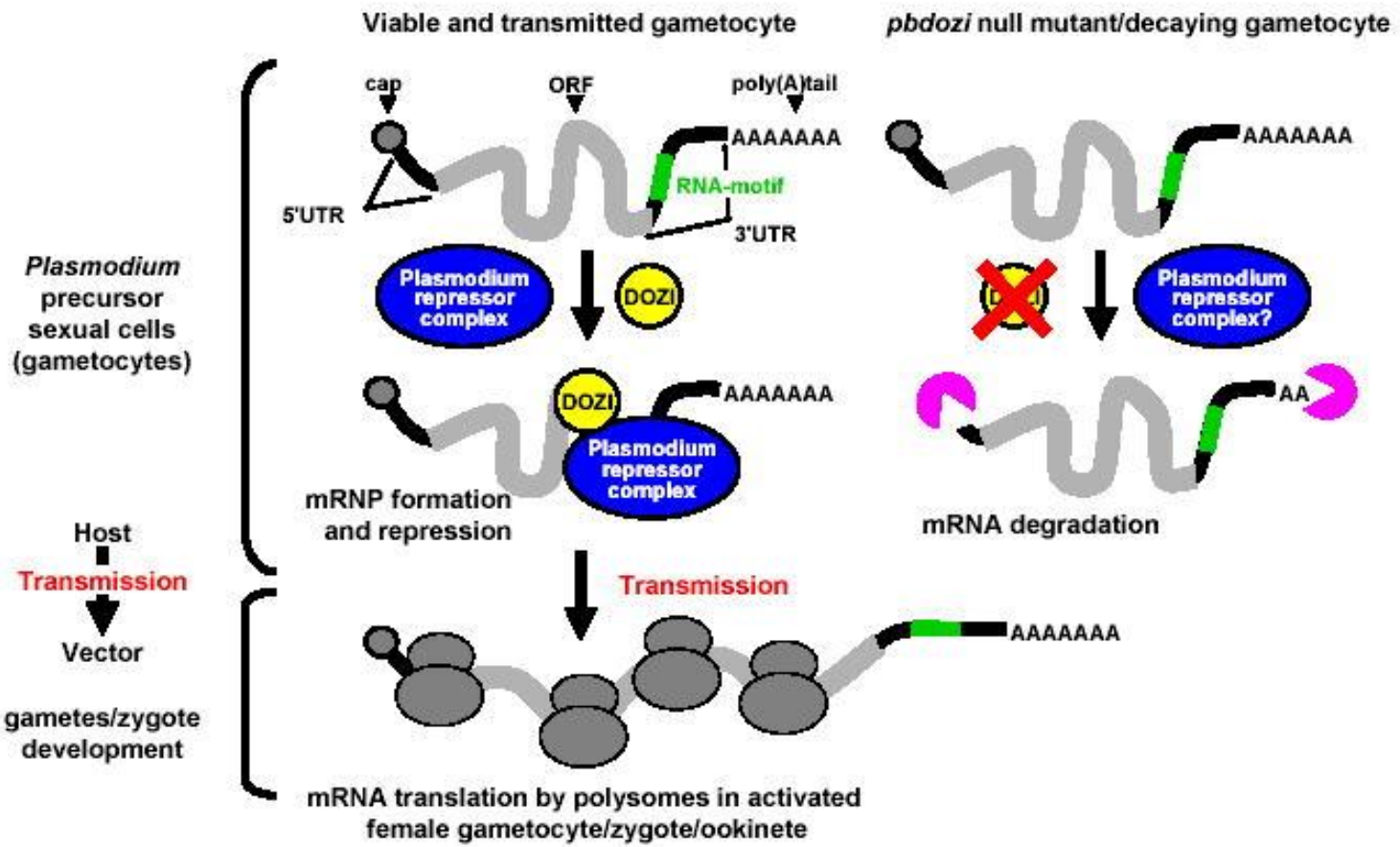
Putative regulatory motif in 3’ UTRs of gametocyte repressed transcripts

- 29 genes identified with 47bp degenerate motif (E value <10-5).
- 22 have clear orthologs in *P. falciparum.*
- 16 Pf orthologs are upregulated in Gametocytes.
- Only 2 identified in Pf. gametocyte proteome.
- Motif was not found in *P. falciparum.*



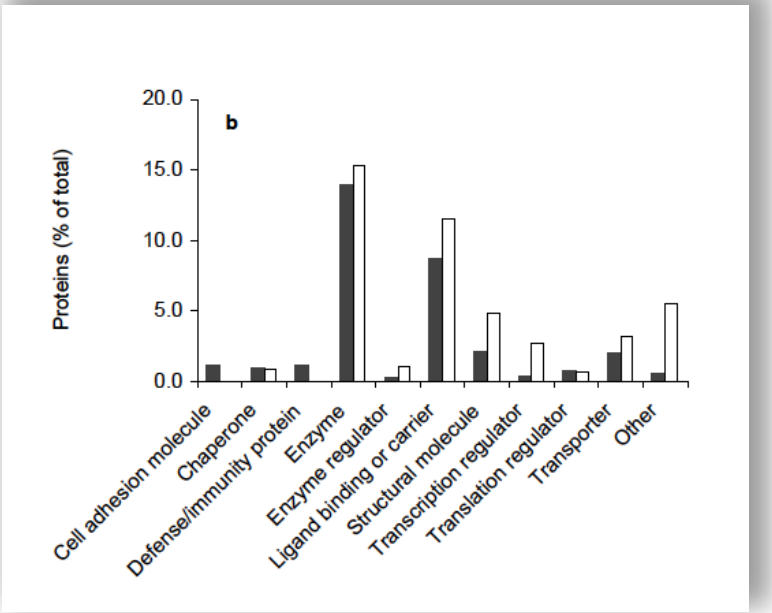
Hall *et al* Science, 2005

Dozi represses transcription in the female gametocyte



Mair , Braks, Garver, Wiegant, **Hall**, et al . 2006 Science. 313:667-9

Why translational control?

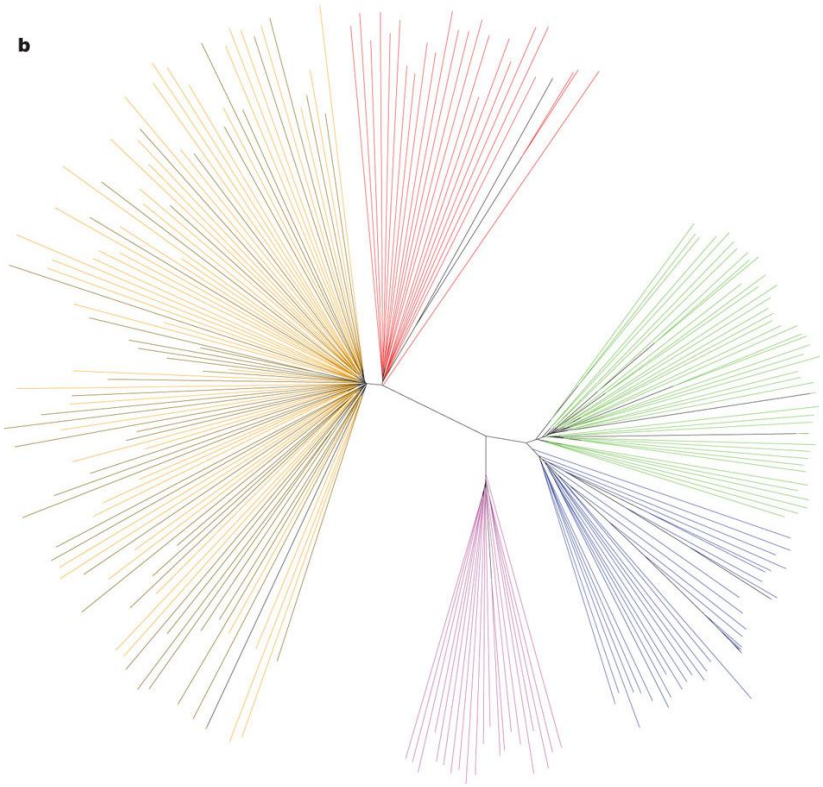


Transcription is ‘relatively’ slow. If your environment changes dramatically this could be a problem

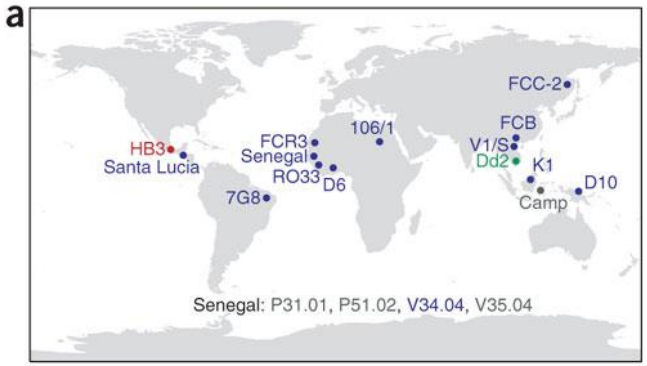
If you know what’s coming you can prepare for it.

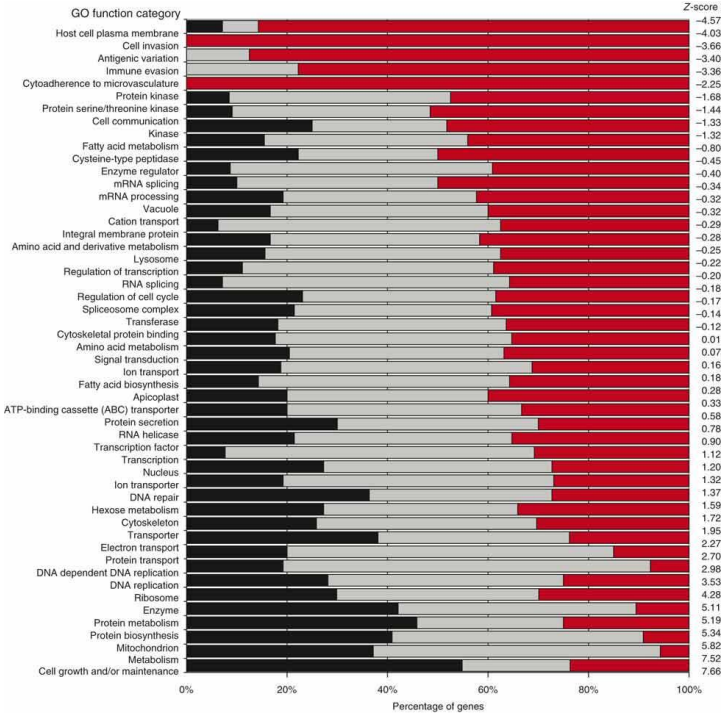
Hence gene regulation by transcriptional control is reduced in P.f

Population genomics in Plasmodium



Within species Plasmodium comparisons





- compared 470 regions 18 strains in total from around the world.
  - PCR amplification and sanger sequence
- Measured total variation in each gene
- Diversity is strongly associated with host interaction

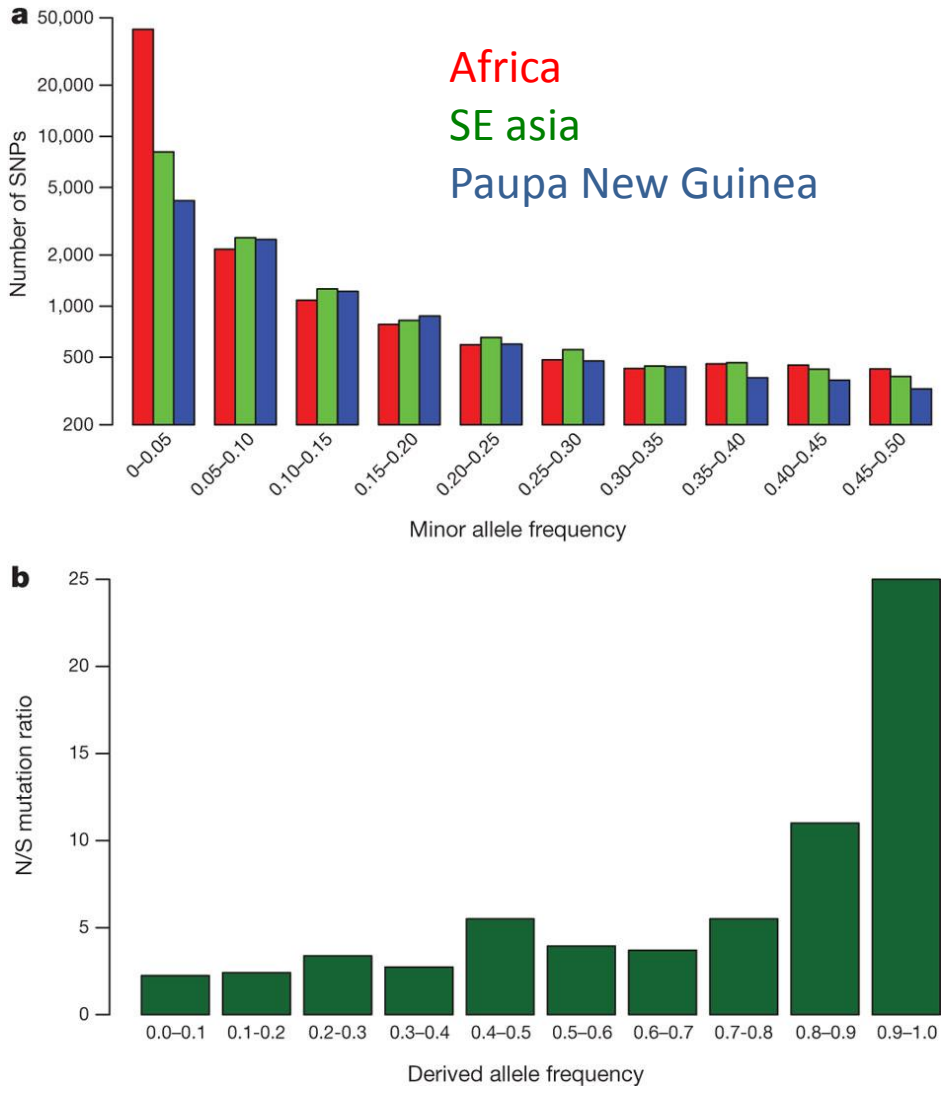
High diversity = Red
Intermediate = Grey
No diversity =Black

Volkman et al (2007) Nature Genetics 39:115

MalariaGen population wide sequecning of malaria

- The MalariaGEN aim is to characterize genetic variation in *P. falciparum* populations.
- Goal is to establish a repository of population genomic data for *P. falciparum* that is of direct and immediate benefit to ongoing scientific research and ultimately to malaria control.
- 1685 samples from 25 separate locations in 17 countries
- Map haplotype frequencies for key mutations associated with drug resistance and other important phenotypes in *P. falciparum*

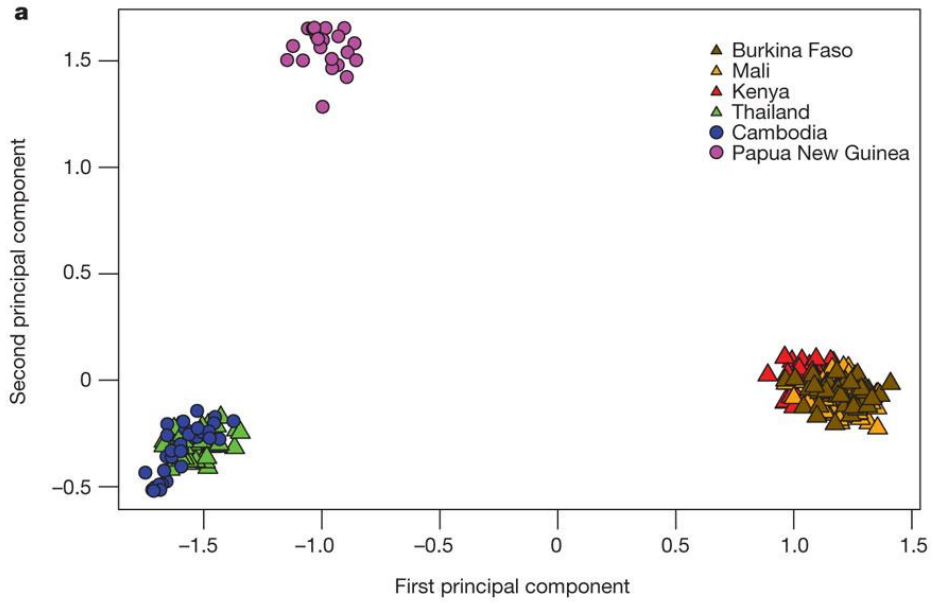
Allele frequency spectrum of SNPs genotyped in this study.

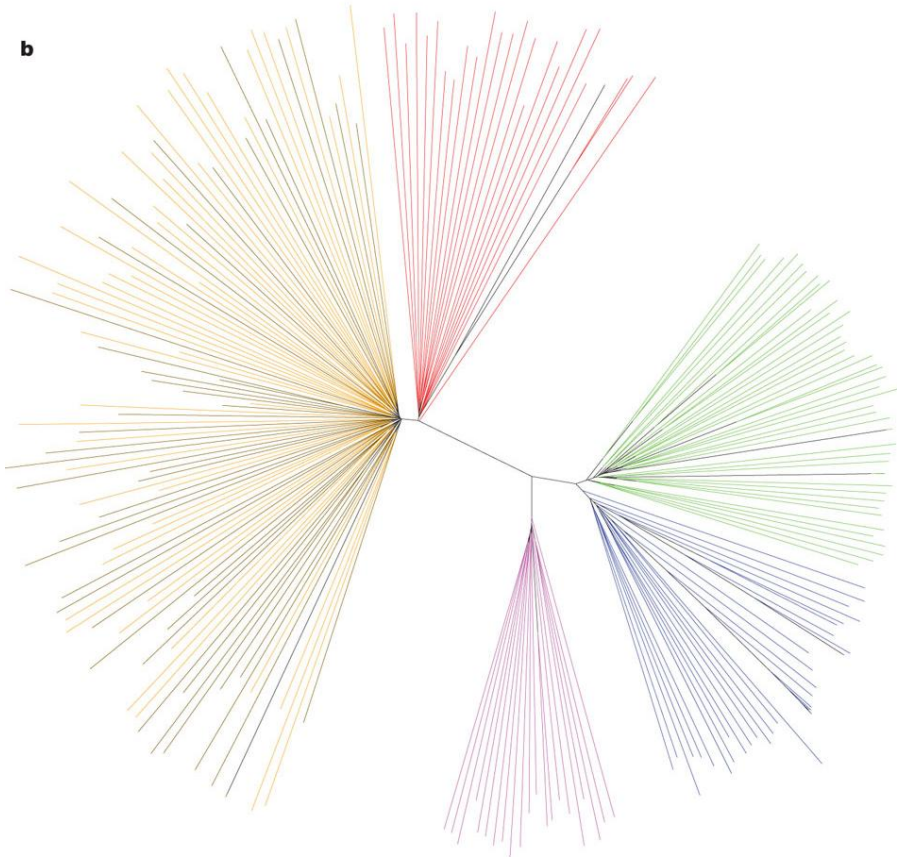


- Minor alleles dominate
  - Consistent with recent expansion
- More low-freq alleles in Africa
  - African origin
- Derived alleles associated with NS mutation
  - Directional selection

Manske (2012) Nature **487**:375

Genetic distance matrix between the 227 samples analysed



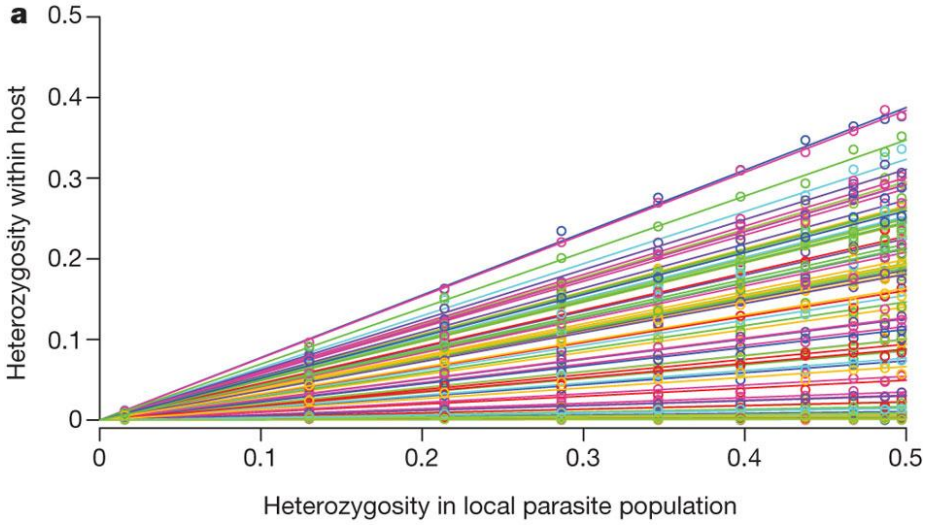


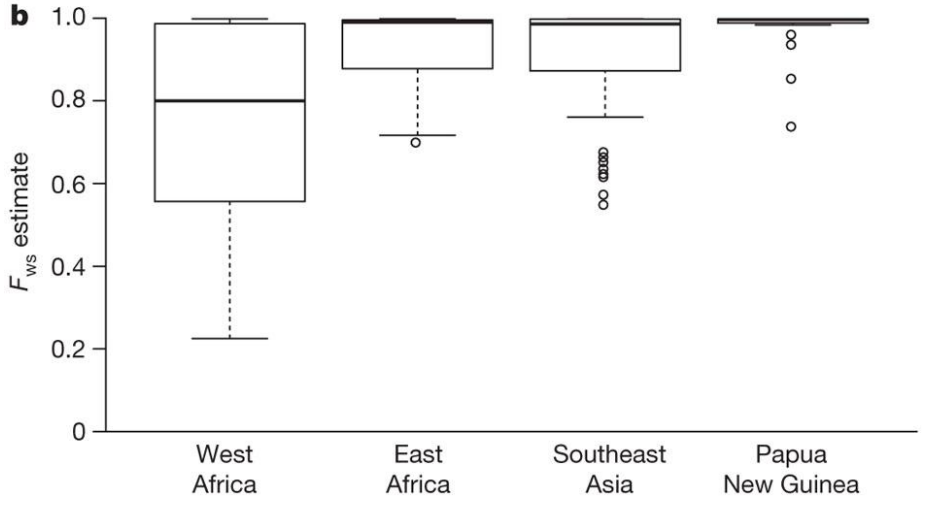
Large separation between continents

Separation of populations in areas of low transmission

Manske (2012) Nature **487**:375

Quantification of within-host diversity





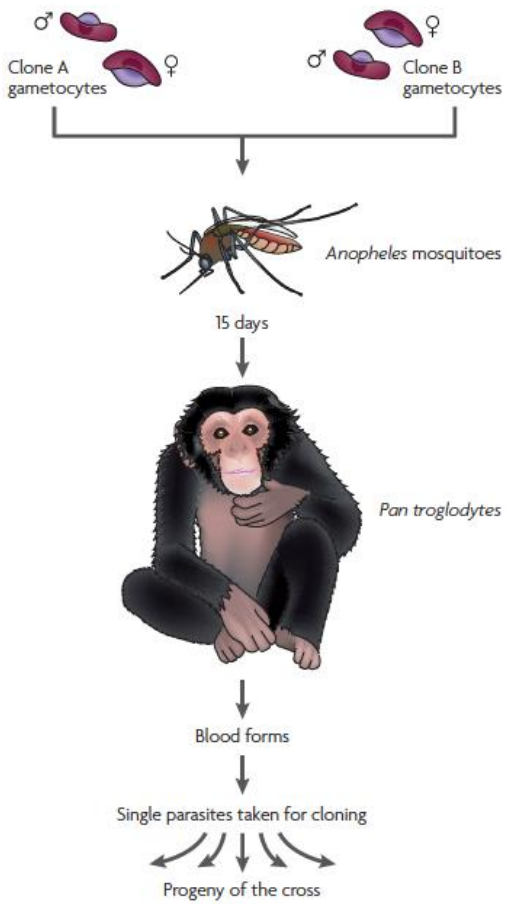
- Inbreeding coefficient (F_{ws}) can be estimated based on heterozygositiy within a host.
- F_{ws} is probably related to transmission rates and human population distribution
- Inbreeding can lead to drug resistance.

Manske (2012) Nature **487**:375

Identifying drug resistance loci

- Traditional genetic methods
- Allele selection
- Genome wide studies for selective sweeps

Genetic crosses for identifying drug resistance loci

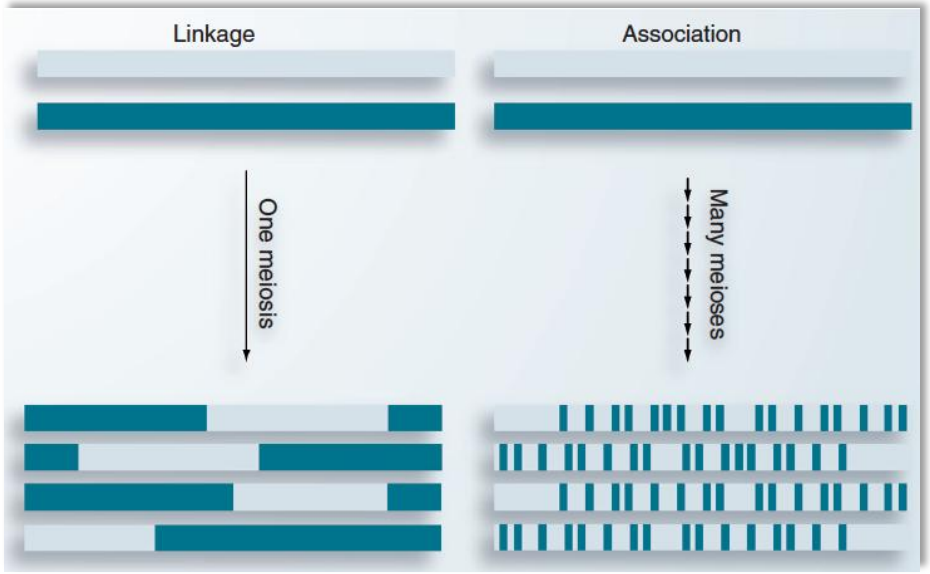


- *P. falciparum* is very host specific. Genetic crosses cant be carried out in vitro
- Crosses require use of higher primates (Chimpanzees) and infection using live mosquitos.
- To date there have been 2 crosses of *P. falciparum*, published for identify drug resistance loci.

Anderson et al Pharmacogenomics (2011) 12(1), 59–85

Traditional methods for identifying drug resistance

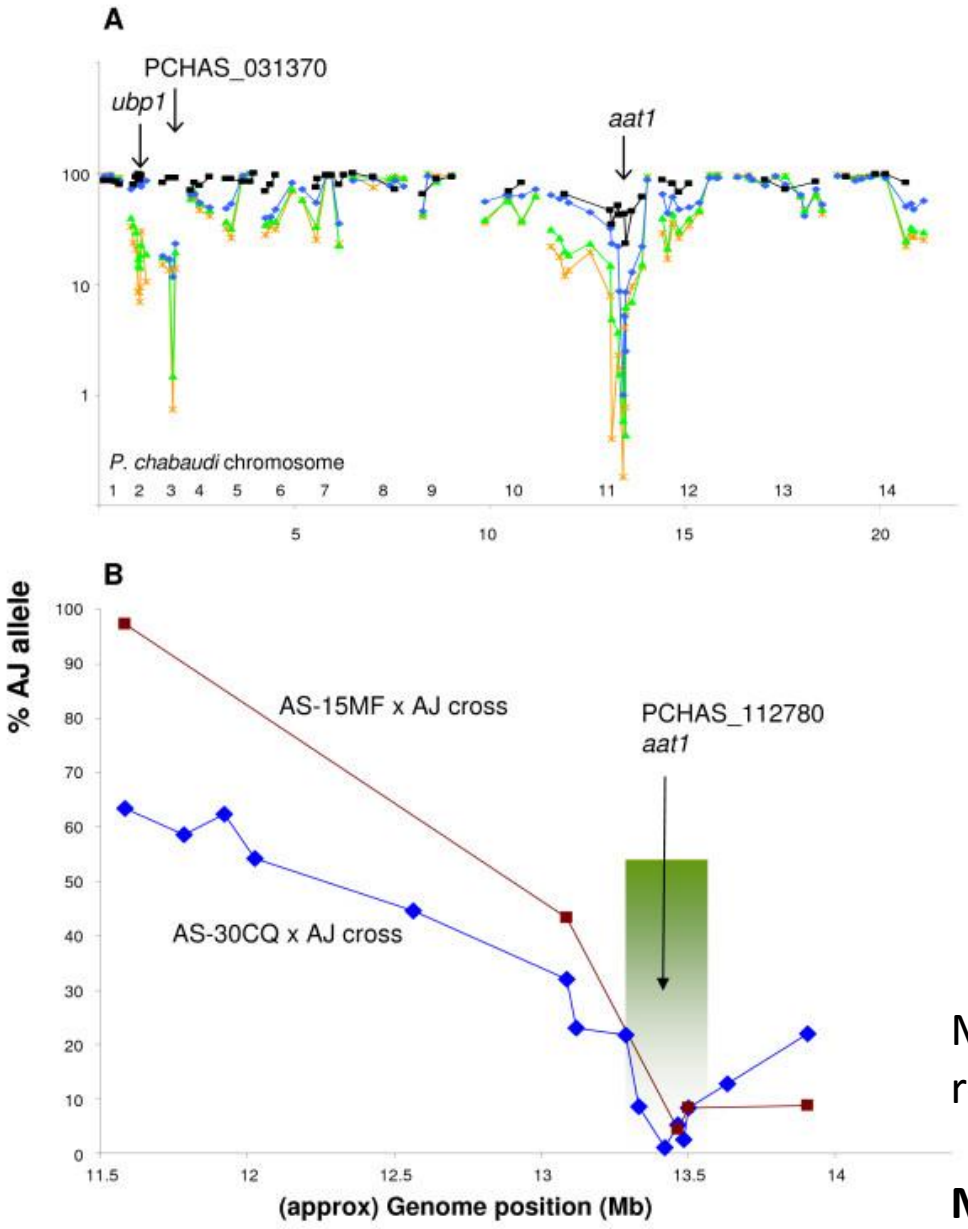
- The recombination rate in *P. falciparum* is high (1 cM = 17 kb) which is approximately 50-times greater than in humans (1 cM = 900 kb), and 15-times greater than *Drosophila* (1 cM = 250 kb).
- This means that mapping an provide relatively good resolution with few markers.



Anderson et al Pharmacogenomics (2011) 12(1), 59–85

- Take 2 strains with different phenotype
- Cross in mosquito and infect vertebrate host
- Clone and phenotype progeny
- Identify markers associated with the phenontype

Allele selection



Allele selection has been used extensively in mouse model, *P. chabaudii*

However the loci that are important in rodent parasites are not always the same as those important in human parasites.

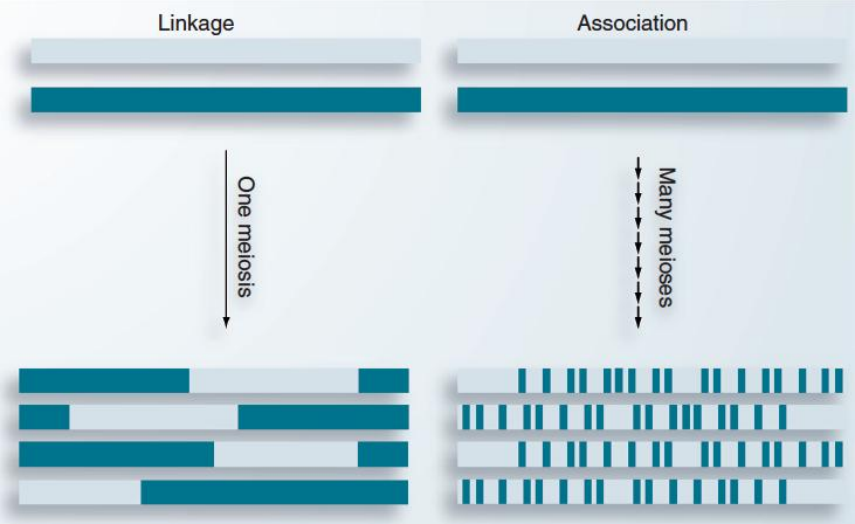
Multiple alleles contribute to chloroquine resistance in P. chabaudii

Modrzynska et al. *BMC Genomics* 2012, **13**:106

GWAS analysis

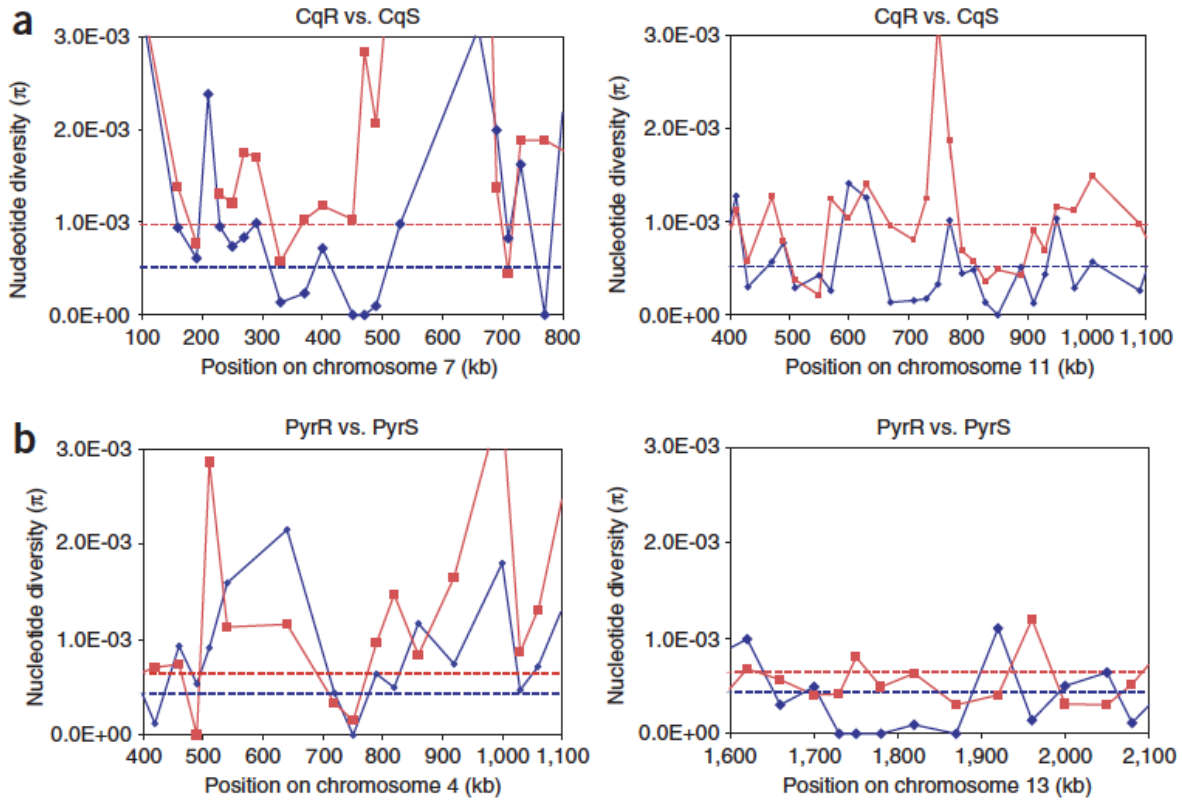
- Alleles under recent selection (i.e from drugs) are often recently derived (new)
- Hence they have not been broken up by recombination
- These alleles are therefore often on large blocks and can be identified in populations as such.

GWAS analysis



- In natural populations haplotypes are much smaller compared to in vitro crosses
- More markers are required to find linkage
- Also there is more variations so more individuals must be analysed to find a significant association

Long range LD and low diversity are a marker for drug resistance



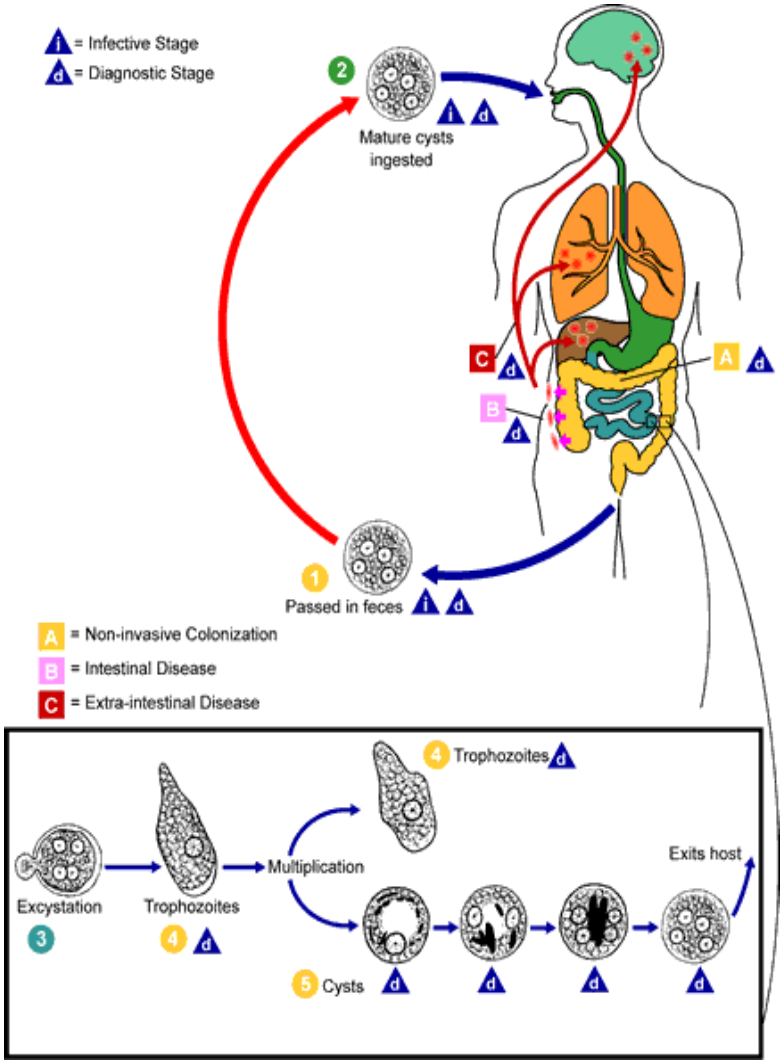
Resistant lines in Blue have less diversity at DR loci than ancestral allele in red

Volkman et al (2007) Nature Genetics 39:115

UNIVERSITY OF LIVERPOOL

Genetics from scratch in *Entamoeba histolytica*

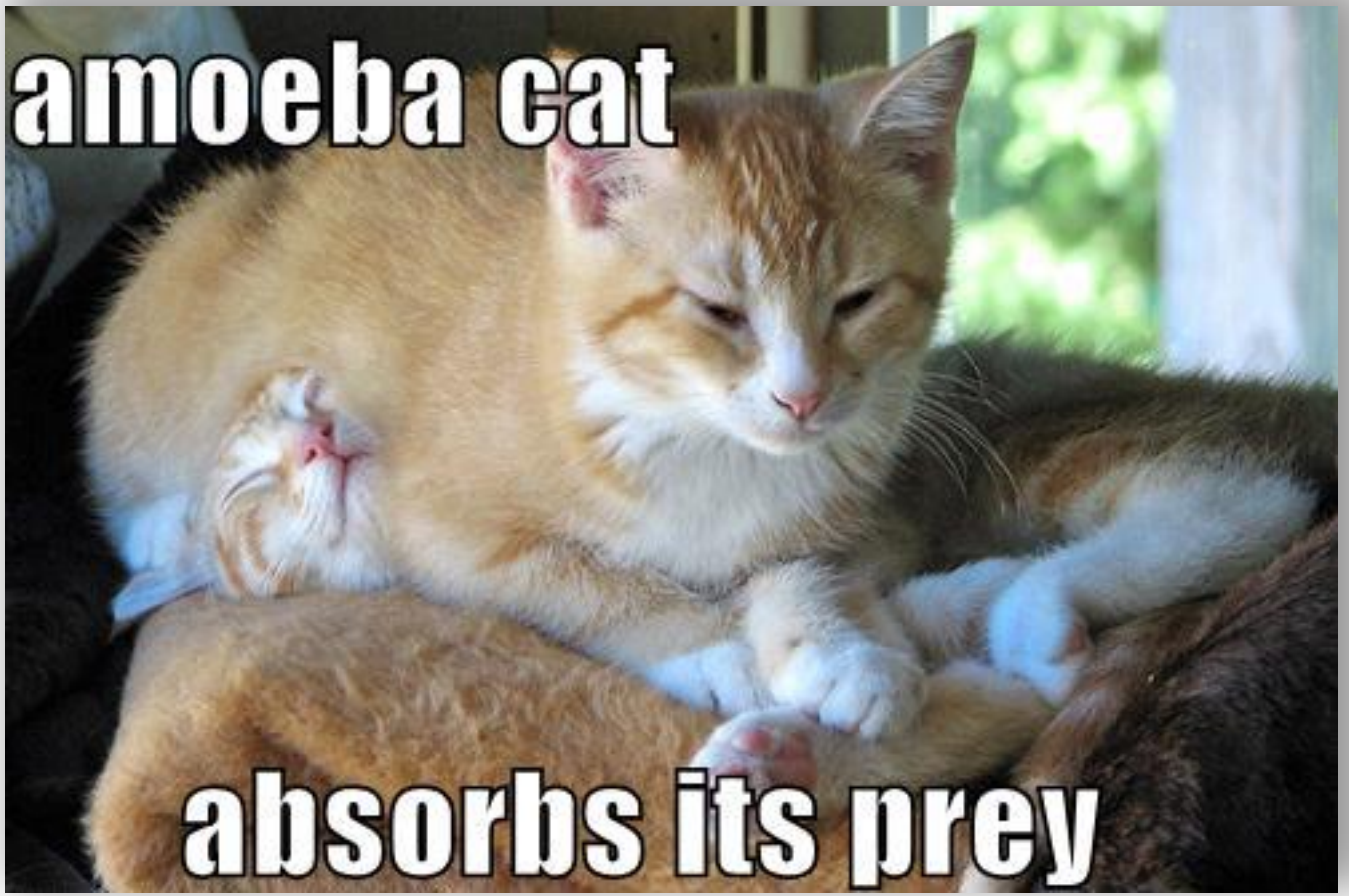
Entamoeba histolytica



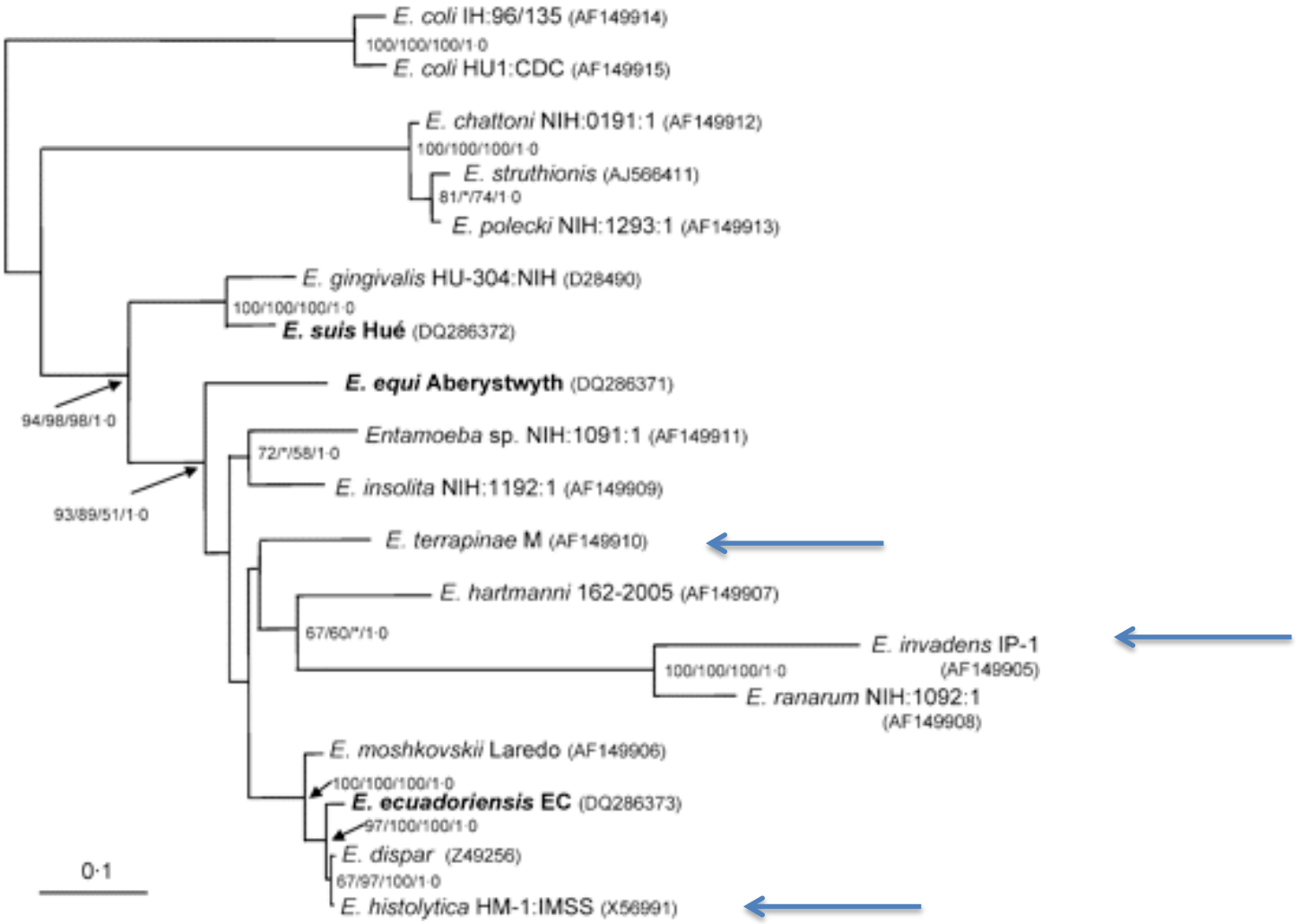
- Transmission is by fecal oral route
- Cysts in contaminated water or food enter gut
- Excystation occurs in the small bowel to form motile trophozoites.
- Encystation of trophozoites produce new cysts that are passed in stools.
- Colitis occurs when parasites invade the gut wall
- Parasites entering the blood stream can invade the liver causing liver abscesses.
- 40-50 million develop invasive disease annually (WHO, 1985)
- 40-110,000 deaths annually
- 80-90% infected people asymptomatic

Entamoeba

http://www.medicine.virginia.edu/


amoeba cat
absorbs its prey

Entamoeba spp are a diverse genus of commensals (and some parasites)



Clark et al (2006) Int J Syst Evol Microbiol **56**, 2235-2239

Findings from *E. histolytica* genome

- **Significant loss/degradation of variety of functions/organelles as a result of parasitic lifestyle in nutrient rich environment**
- **Concomitant expansion in specific domain families related to vesicle trafficking/phagocytosis & sensing of its environment**
- **Some unusual metabolic pathways may be present based on absence of synthesis pathways for key metabolites .**
- **Lateral gene transfer of genes related to various aspects of metabolism into the genome.**
- **Unusual or bacterial like processes may serve as good therapeutic targets as they are not found in the host.**

letters to nature

The genome of the protist parasite *Entamoeba histolytica*

Brendan Loftus[1], Iain Anderson[1], Rob Davies[2], U. Cecilia M. Alsmark[3], John Samuelson[1], Paolo Amedeo[1], Paola Roncaglia[1], Matt Berriman[2], Robert P. Hirt[3], Barbara J. Mann[4], Tomo Nozaki[5], Bernard Suh[1], Mihai Pop[1], Michael Duchene[7], John Ackers[8], Egbert Tannich[9], Matthias Leippe[10], Margit Hofer[7], Iris Bruchhaus[6], Ute Willhoeft[9], Alok Bhattacharya[11], Tracey Chillingworth[2], Carol Churcher[2], Zahra Hance[2], Barbara Harris[2], David Harris[2], Kay Jagels[2], Sharon Moule[2], Karen Mungall[2], Doug Ormond[2], Rob Squares[2], Sally Whitehead[2], Michael A. Quail[2], Ester Rabbinowitsch[2], Halina Norbertczak[2], Claire Price[2], Zheng Wang[1], Nancy Guillén[12], Carol Gilchrist[5], Suzanne E. Stroup[13], Sudha Bhattacharya[11], Anuradha Lohia[13], Peter G. Foster[14], Thomas Sicheritz-Ponten[15], Christian Weber[12], Upinder Singh[16], Chandrama Mukherjee[13], Najib M. El-Sayed[1], William A. Petri Jr[5], C. Graham Clark[3], T. Martin Embley[3], Bart Barrell[2], Claire M. Fraser[1] & Neil Hall[2]*

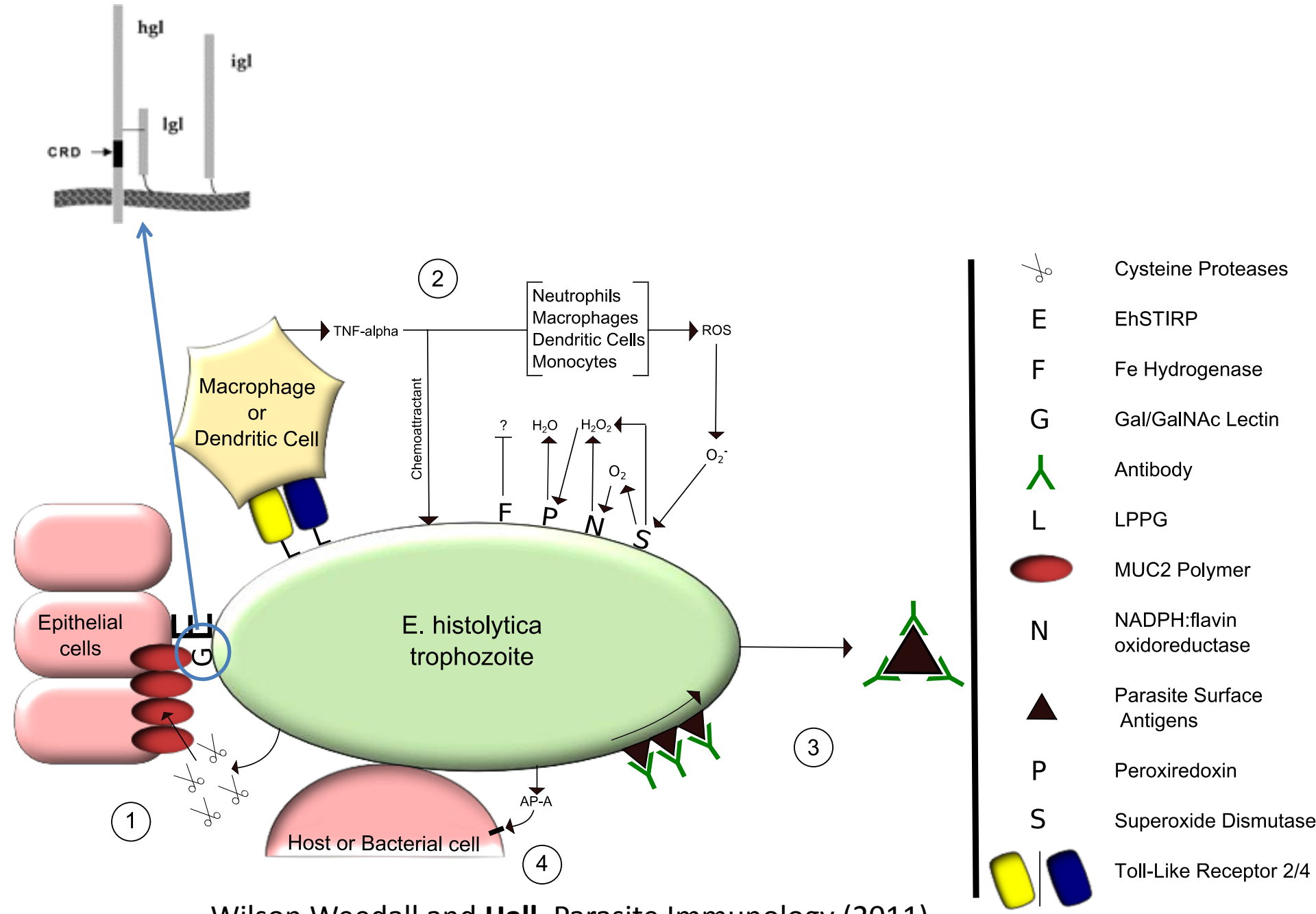
[1]*TIGR, 9712 Medical Center Drive, Rockville, Maryland 20850, USA*
[2]*The Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
[3]*School of Biology, University of Newcastle, King George VI Building, Newcastle upon Tyne NE1 7RU, UK*
[4]*Department of Molecular and Cell Biology, Boston University Goldman School of Dental Medicine, 715 Albany Street, Boston, Massachusetts 02118, USA*
[5]*Departments of Internal Medicine & Microbiology, University of Virginia, Charlottesville, Virginia 22908, USA*
[6]*Department of Parasitology, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8640, Japan*
[7]*Division of Specific Prophylaxis and Tropical Medicine, Center for Physiology and Pathophysiology, Medical University of Vienna, Kinderspitalgasse 15, A-1095 Vienna, Austria*
[8]*Department of Infectious and Tropical Diseases, London School of Hygiene and*

gene families, including those associated with virulence. Additional genome features include an abundance of tandemly repeated transfer-RNA-containing arrays, which may have a structural function in the genome. Analysis of the genome provides new insights into the workings and genome evolution of a major human pathogen.

Genome analysis was carried out on a 12.5-fold coverage genome assembly consisting of 23,751,783 base pairs (bp) distributed among 888 scaffolds. The 9,938 predicted genes average 1.17 kilobases (kb) in size and comprise 49% of the genome. One-quarter of *E. histolytica* genes are predicted to contain introns, with 6% of genes containing multiple introns. No homologues could be identified for a third of predicted proteins (31.8%) from the public databases (see Methods). *E. histolytica* chromosomes do not condense, and the uncertainty surrounding its ploidy and the extensive length variability observed between homologous chromosomes from different isolates makes the exact chromosome number difficult to determine. The chromosome size variation observed may be due to expansion and contraction of subtelomeric repeats, as in other protists[2,3], and it is tempting to speculate that in *E. histolytica* these regions consist of tRNA-containing arrays. Comprising almost 10% of the sequence reads, 25 types of long tandem array, each containing between one and five tRNA types per repeat unit, could be identified from the genome data. The full complement of tRNAs required for translation has been identified, and all but four of the tRNA genes are encoded exclusively in arrays. These unique tRNA gene arrays are thus predicted to be functional as well as potentially fulfilling a structural role in the genome. No association could be determined between codon usage and the relative copy numbers of their cognate tRNA species.

The metabolism of *E. histolytica* seems to have been shaped by secondary gene loss and lateral gene transfer (LGT), primarily from bacterial lineages (Fig. 1). *E. histolytica* is an obligate fermenter, using bacterial-like fermentation enzymes and lacking proteins of

Loftus et al (2005) Nature **433**:865



Wilson Weedall and **Hall**. Parasite Immunology (2011)

Genetics of *E. histolytica*

- Disease is only occurs in a minority of cases of *E. histolytica*
- Some strains are more virulent than others in model infections (eg Rahman vs HM1)
- We hypothesise that there is an underlying genetic cause for this
- Unfortunately as there are no microsatellites in the *E. histolytica* genome no genetic studies have been performed

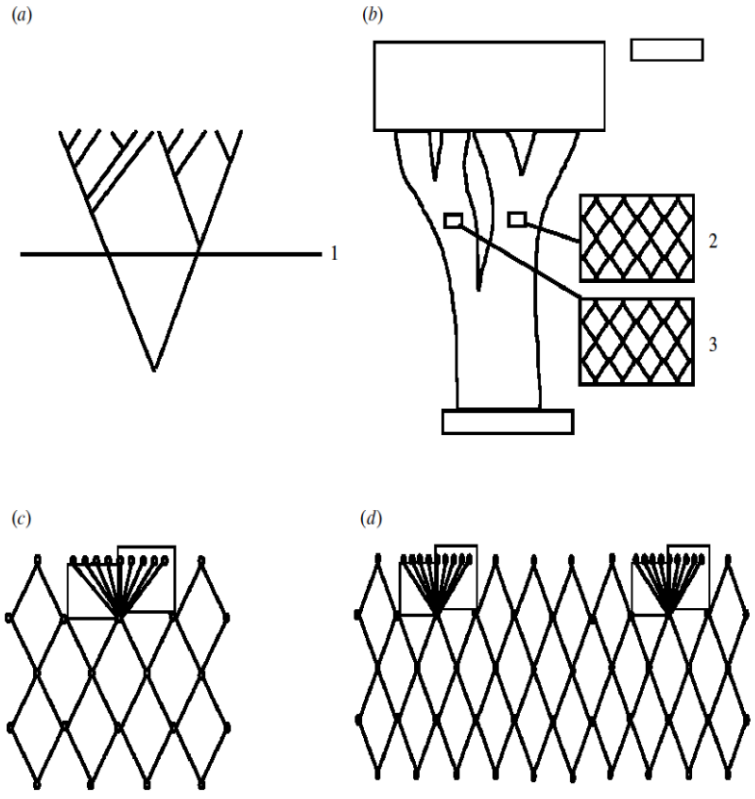
Why sex is important

- Recombination of genes
- New variants
- Allows establishment into new niches
- Variation helps adapt to immune system of host
- Spread of important traits .. drug resistance/virulence

Lots of Sex vs Occasional Sex vs No Sex

- Some parasites (Plasmodium) have meiosis every life cycle
- Others only replicate clonally (*Leishmania major*)
- Others have clonal epidemics with occasional recombination between strains (*T. brucei*)
- Others we are not sure about….

Maynard Smith et al (1993)



A-Clonal
B-Panmictic
C & D - Epidemic

Genome re-sequencing

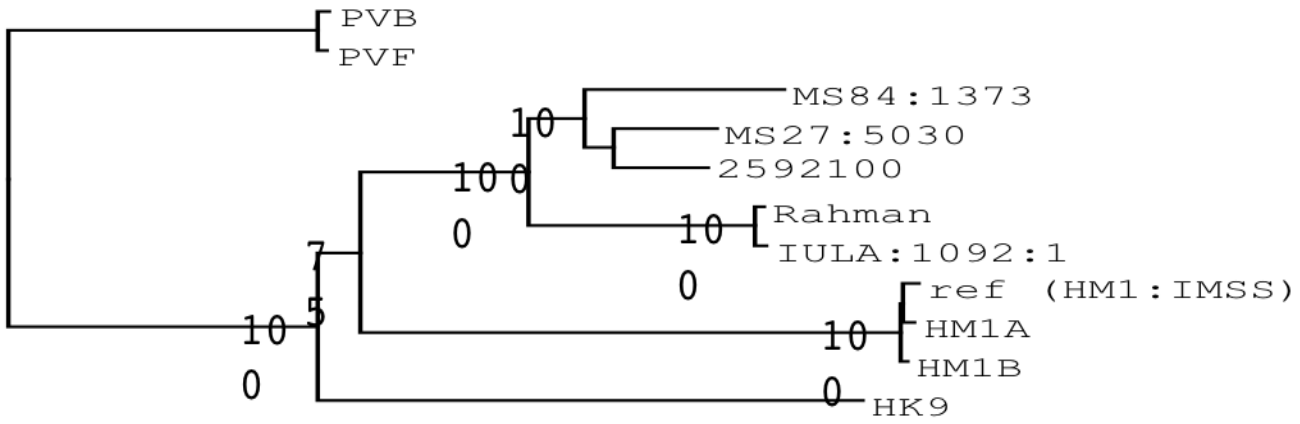
Weedall et al Genome Biol. 2012 13(5):R38

8 axenic strains sequenced (provided by C. Graham Clark, LSHTM)

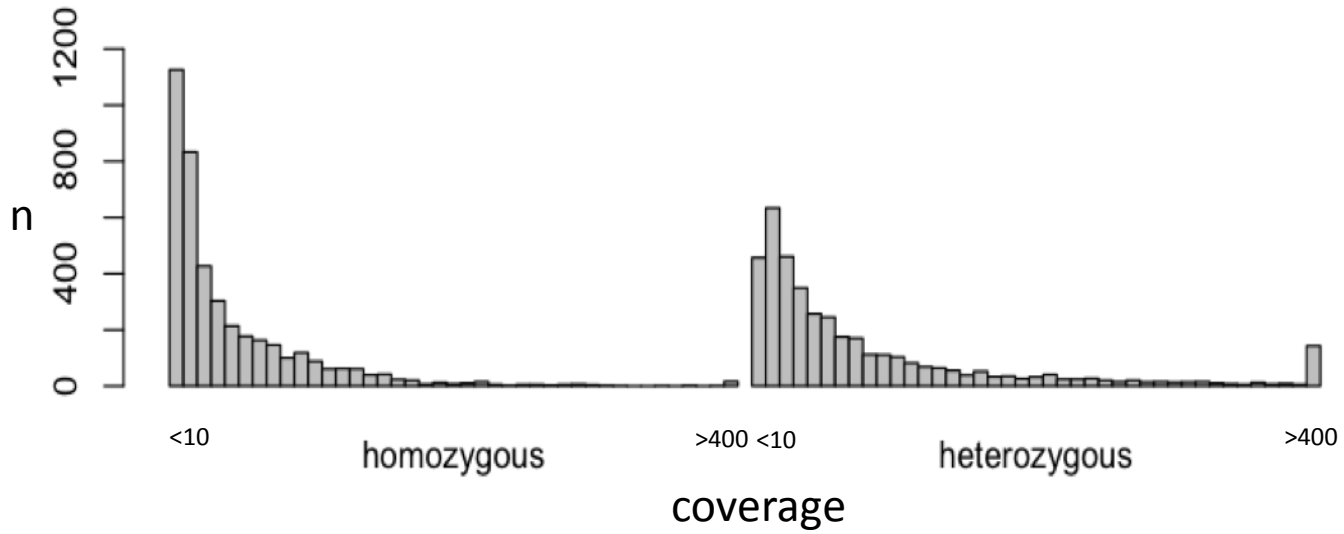
strain	origin	date	phenotype
Rahman	UK*	1964	asymptomatic
2592100	Bangladesh	2005	intestinal amoebiasis
PVB	Itlay** (biopsy)	2007	intestinal amoebiasis
PVF	Italy** (faeces)	2007	intestinal amoebiasis
HK9	Korea	1951	intestinal amoebiasis
IULA:1092:1	Venezuela	1992	intestinal amoebiasis
MS84:1373	Bangladesh	2006	asymptomatic
MS27:5030	Bangladesh	2006	asymptomatic

* Unknown origin (isolated from sailor)

** Unknown origin (possibly Liberia or Colombia)



Strain	reads	c50,c95 cov	SNP-able sites	SNP	hom	het	divergence(hom only)
HM-1A	13743406	35,141	10012951	2217	229	1988	0.22(0.02)
HM-1B	9586924	26,95	9819882	1995	220	1775	0.20(0.02)
Rahman	19498380	32,198	9817503	6889	3767	3122	0.70(0.38)
2592100	13560609	26,127	10025805	6788	3128	3660	0.68(0.31)
PVBM08B	17627870	36,172	10335217	7999	4225	3774	0.77(0.41)
PVBM08F	8436907	19,65	10253328	6602	3613	2989	0.64(0.35)
IULA:1092:1	19041335	48,155	11934434	10014	4897	5117	0.84(0.41)
HK-9	21193087	41,202	10678584	9155	4428	4727	0.86(0.41)
MS84-1373	21479273	51,209	10308534	8373	4027	4346	0.81(0.39)
MS27-5030	20403218	47,225	8731329	7001	3302	3699	0.80(0.38)



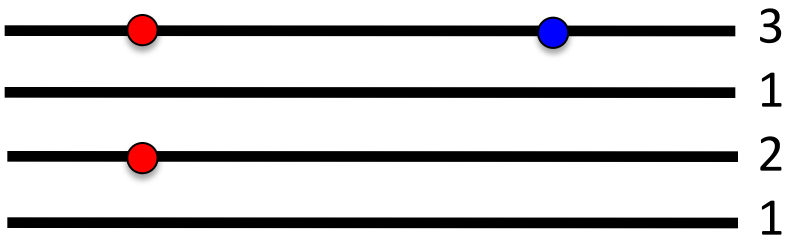
Defined ~4000 'high quality candidate marker' sites = homozygous, coding, in all strains

Weedall et al Genome Biol. 2012 13(5):R38

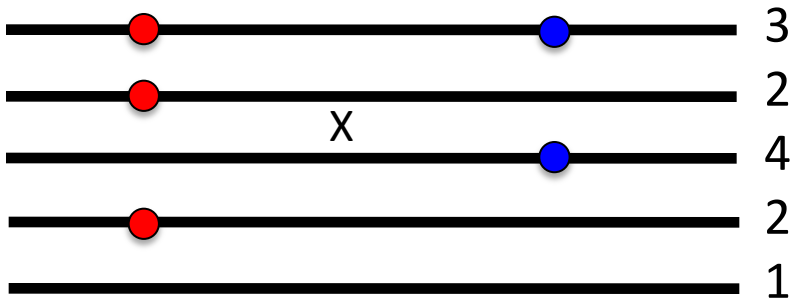
A test for recombination

4-gametes (4-haplotypes) test:

4 haplotypes *must* be due to recombination under the infinite sites model (no recurrent mutation)



- New mutation creates new haplotype

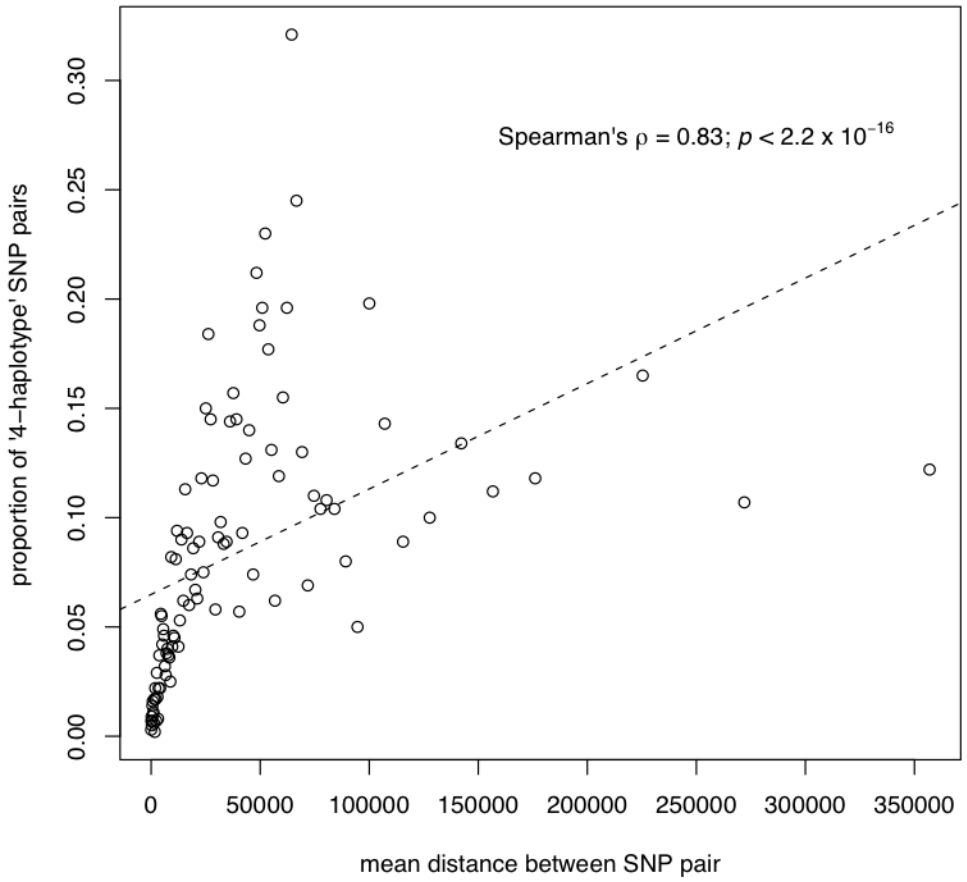


} Recombination between haplotypes 1 and 3 creates new haplotype

Weedall et al Genome Biol. 2012 13(5):R38

A test for recombination

Used ~4000 ‘high quality candidate marker’ sites



average distance between:
2-hap sites = 18,454 bp
3-hap sites = 46,644 bp
4-hap sites = 60,176 bp

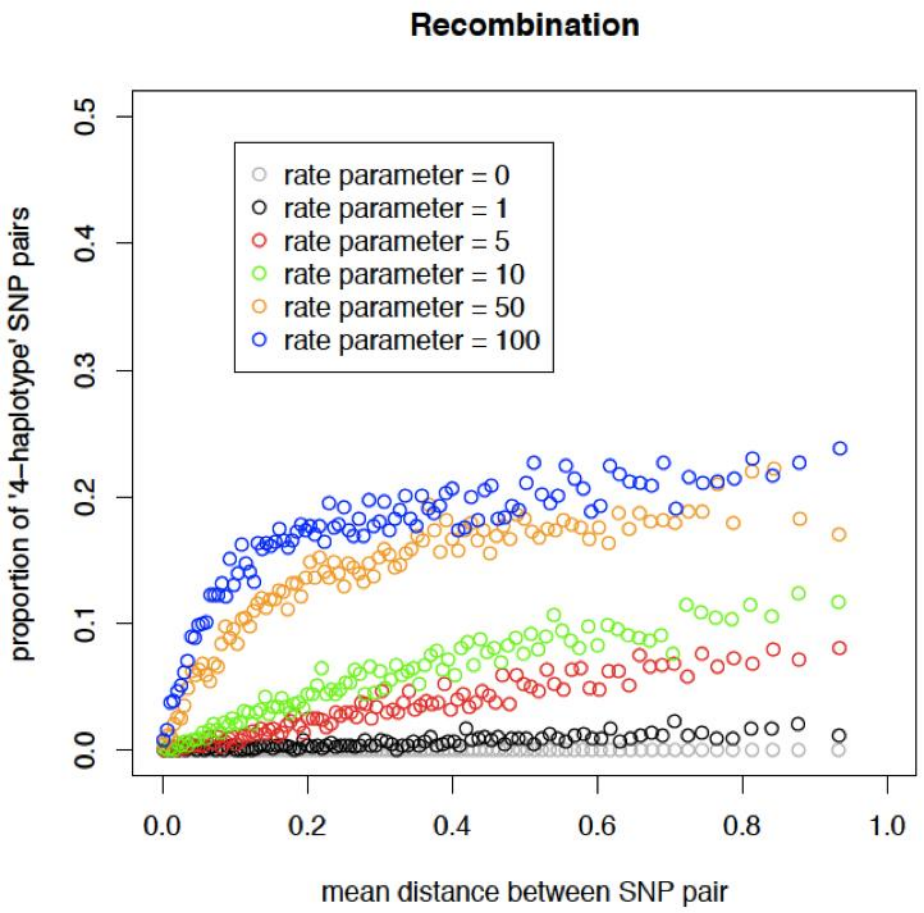
Proportion of site-pairs with 4 haplotypes increases with distance (on same reference scaffold)

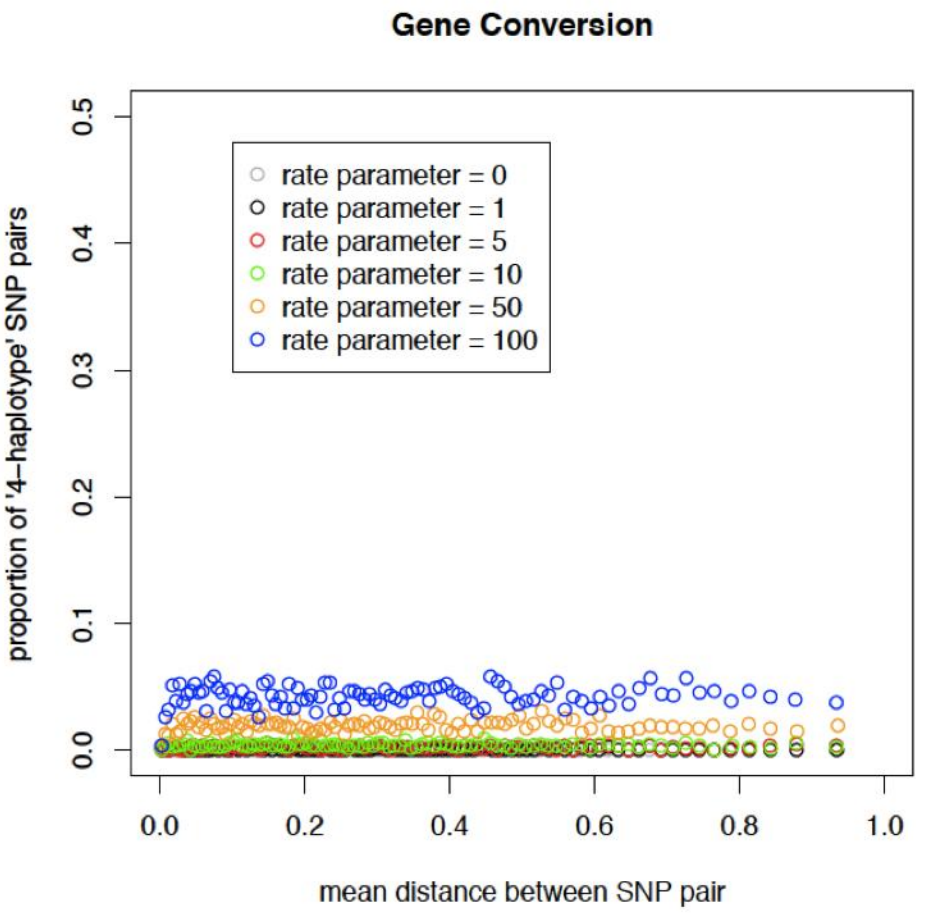
Pattern would be expected given recombination (probability increases with distance)

Suggests historical recombination events

Weedall et al Genome Biol. 2012 13(5):R38

Could we be observing gene conversion?





Recombination was modeled by specifying a recombination parameter 4.Ne.r,

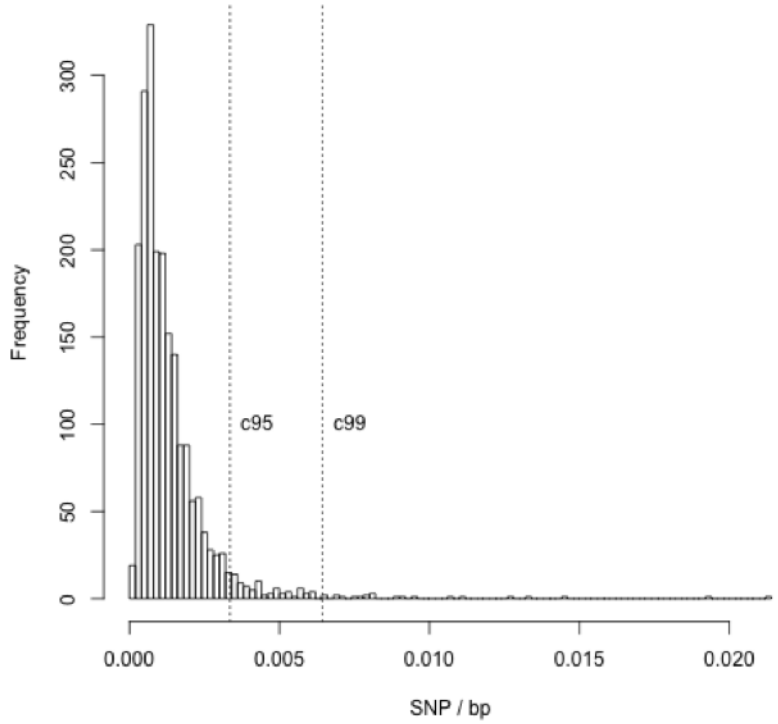
where Ne is the effective population size and r is the per generation probability of a crossover occurring.

Gene conversion was modeled by specifying a parameter 4.Ne.f,

where f in the per generation probability of a gene conversion event in the sequence as well as the length of the 'converted' region.

Sources of variation in *E. histolytica*

- SNP
- Gene loss/gain
- Copy number variation

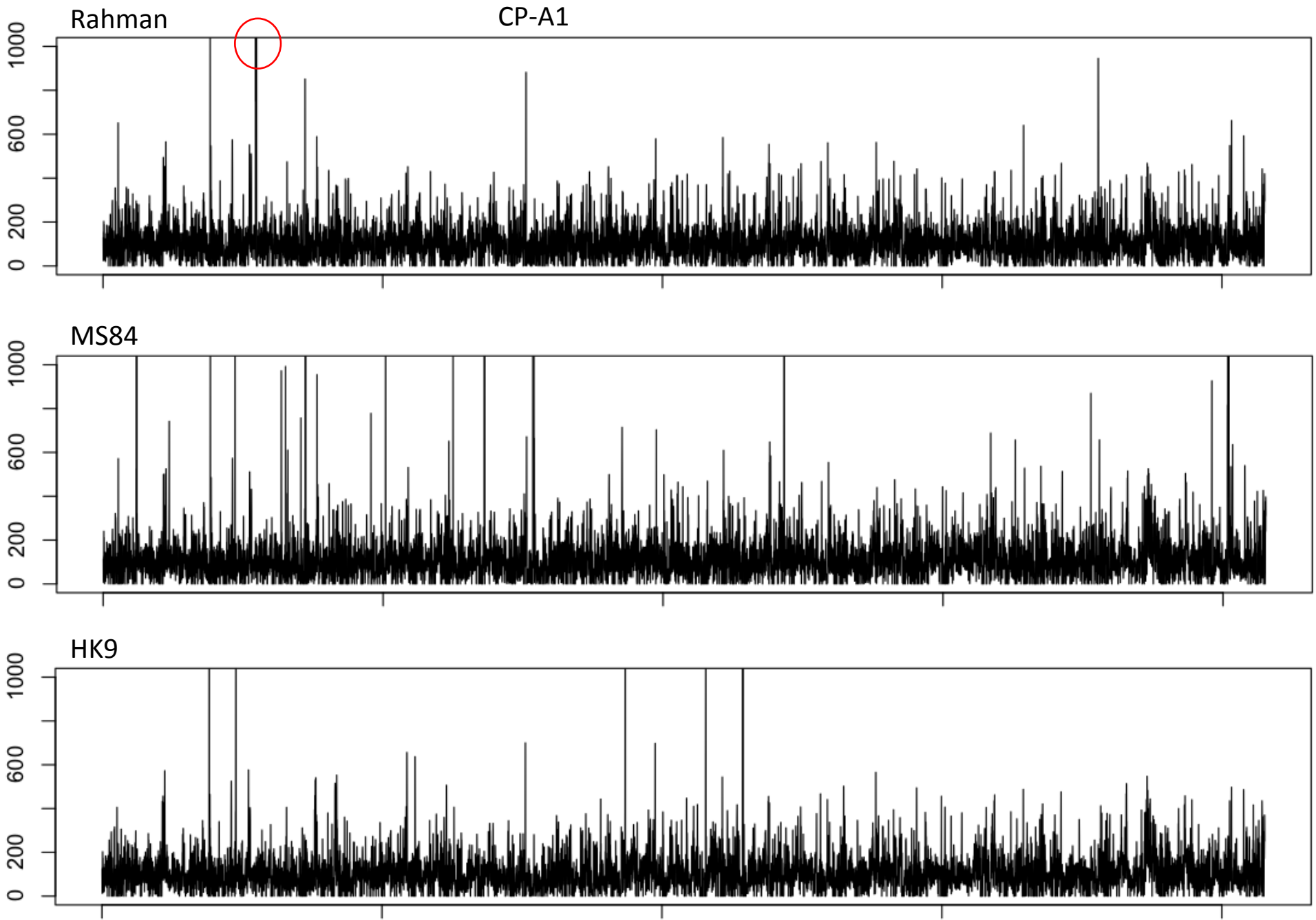


2352 / 8306 genes are polymorphic

~5 % have >= 4 SNP kb^{-1}

~1 % have >= 6 SNP kb^{-1}

Genomic plasticity: gene copy number variation



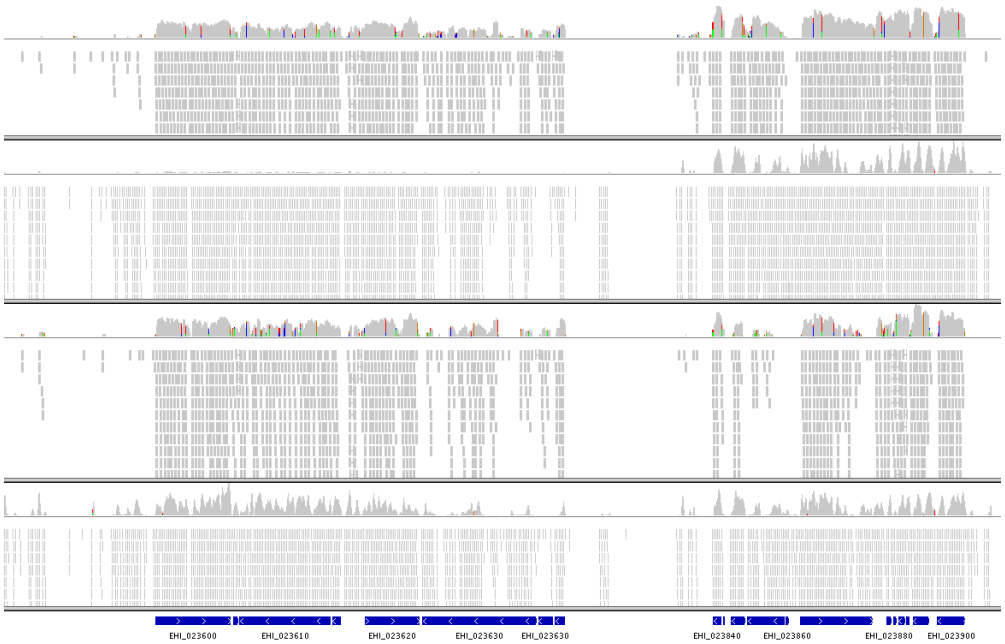
Copy number variation between strains is associated with differential transcription

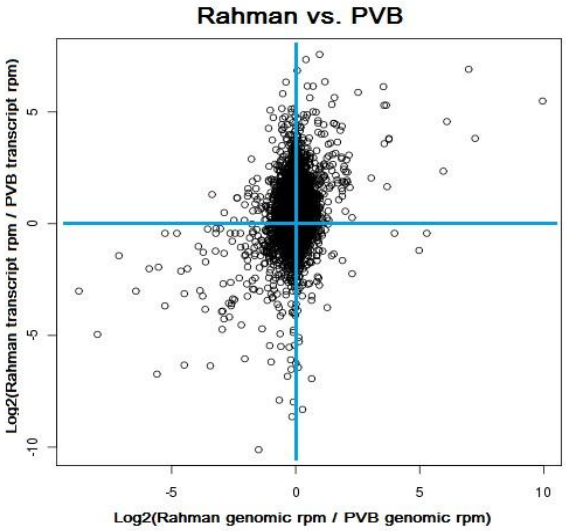
Rahman, transcriptome

Rahman, genome

PVBM08B, transcriptome

PVBM08B, genome





Preativatanyou et al unpublished

Sources of intra-species genetic variation

- Recombination
- Genome plasticity
  - CNV –Gene loss gain
- These processes will affect the emergence and spread of
  - Drug resistance
  - Virulence

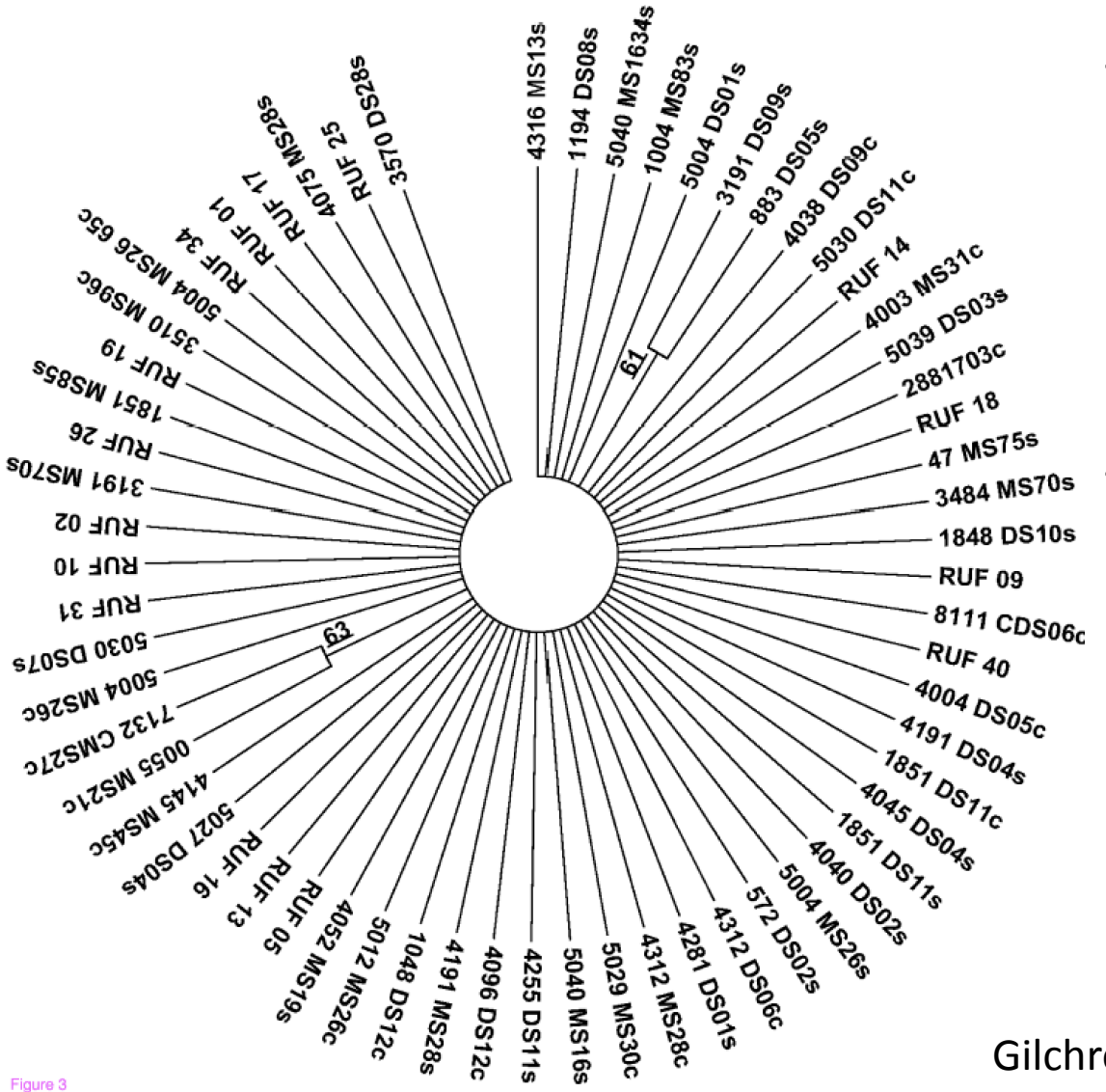
Genotyping of field samples

Gilchrest et al BMC Microbiol. 2012 12:151.

- 19 amebic liver aspirates;
- 26 xenic cultures
  - 14 from asymptomatic infections
  - 12 from diarrheal infections
- 20 *E. histolytica* positive samples from diarrheal stool
- 21 marker loci selected for genotyping by Illumina sequencing.

Rajshahi, Bangladesh

SNP Genotypes from a single population are consistent with frequent recombination

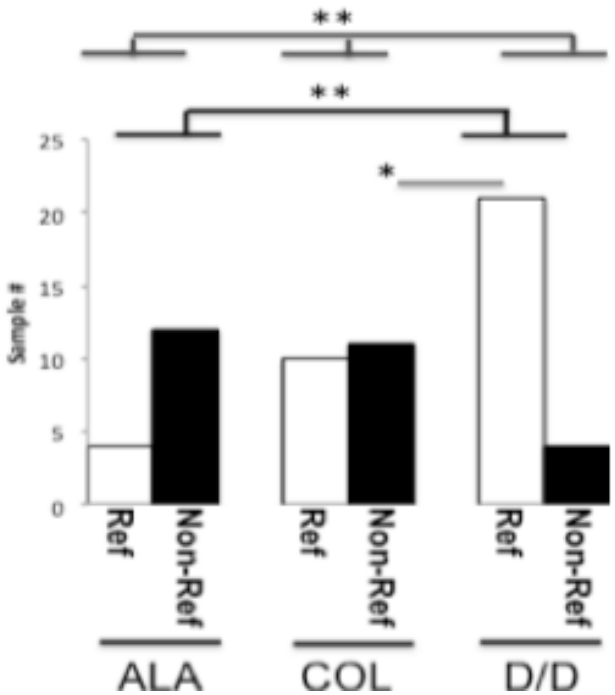


- distribution of the SNP suggest a highly recombining population with few clustered branches.
- Suggesting a lot of recombination between alleles

Gilchrest et al BMC Microbiol. 2012 12:151.

Figure 3

SNPs in the EHI_080100 locus segregate with disease



□ Ref; homozygous Reference allele;
■Non-Ref; homozygous Non-Reference allele

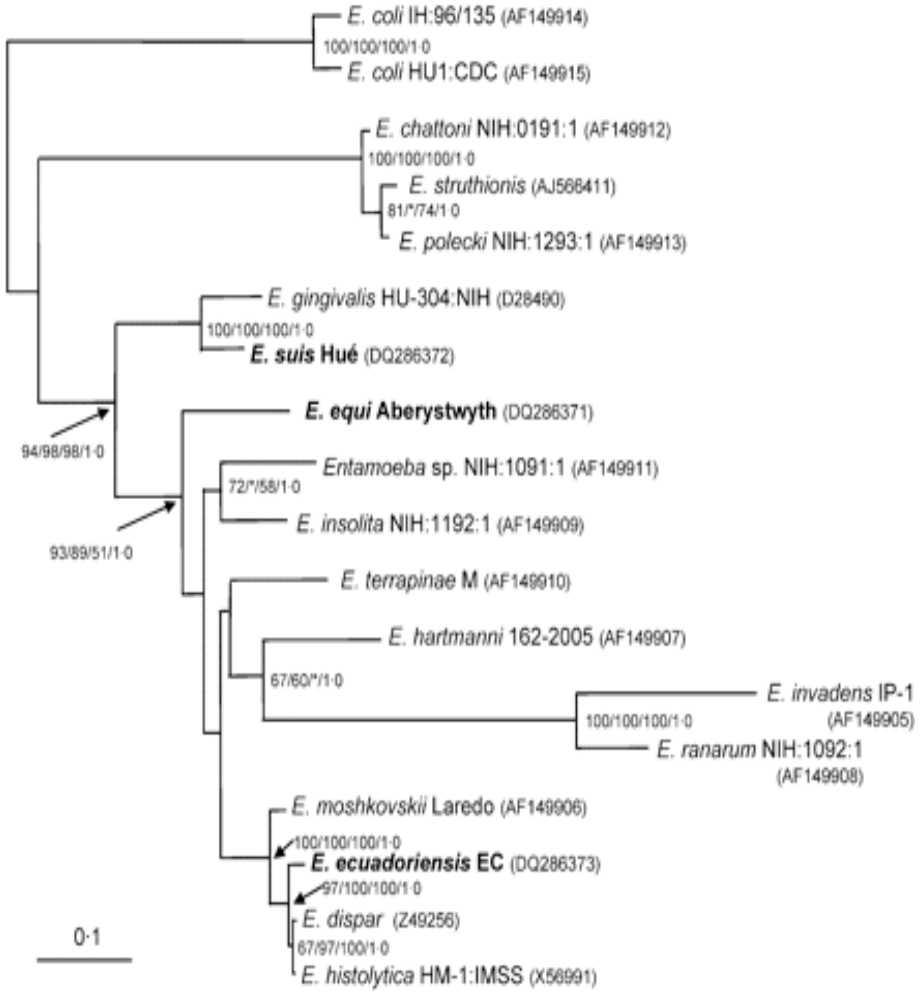
Reference alleles significantly more common in the Diarrhea/Dysentery sample relative to the Colitis and Liver abscess sample

Gilchrest et al BMC Microbiol. 2012 12:151.

If recombination is occurring…
when is it occurring?

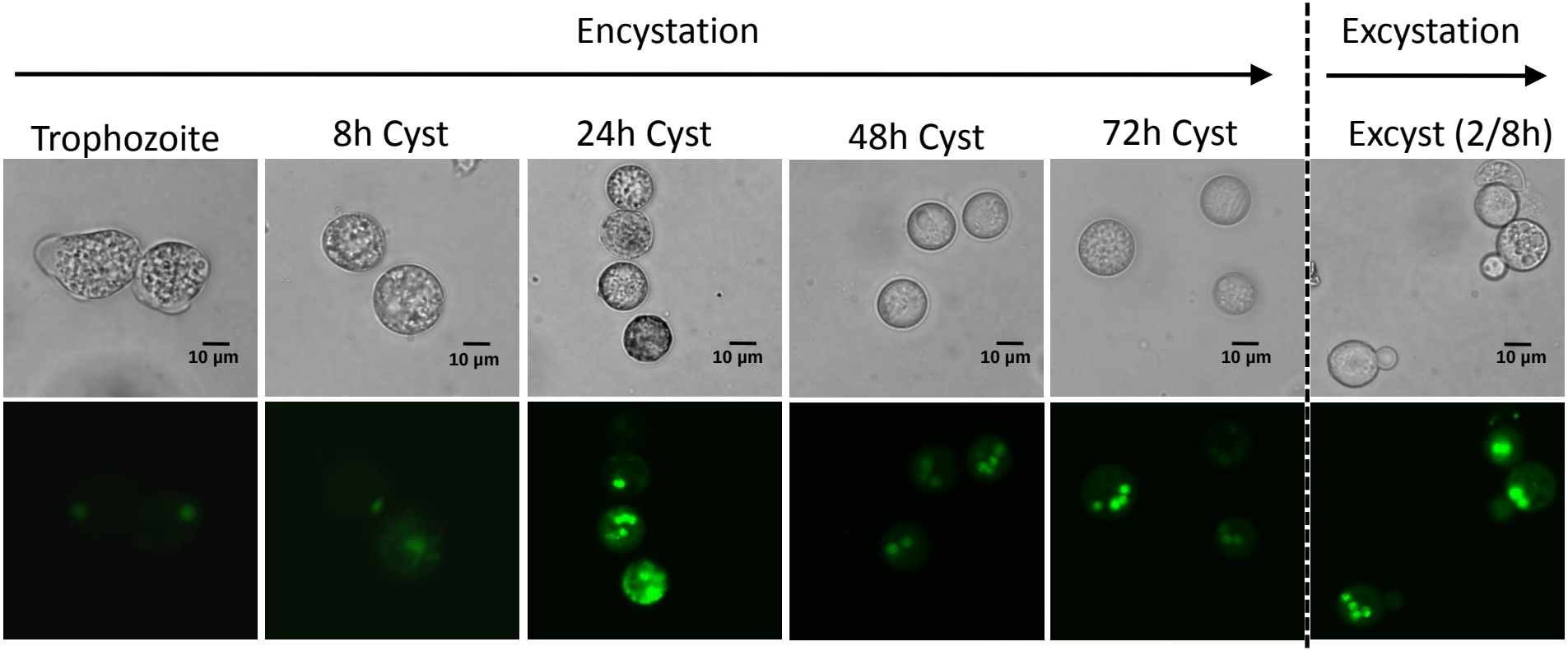
Transcriptome of development in Entamoeba

Ehrenkaufer et al Genome Biol. 2013 14(7):R77



- Using E. invadens as a model for Encystation
- Extract RNA over time-course
- Reads mapped using TopHat
- Expression profiles clustered using STEM
- Differential expression using CuffDiff

Changes in cell morphology during induced encystation and excystation



Intra-timepoint correlation:

0.76	0.96	0.80	0.82	0.50	0.98/0.98

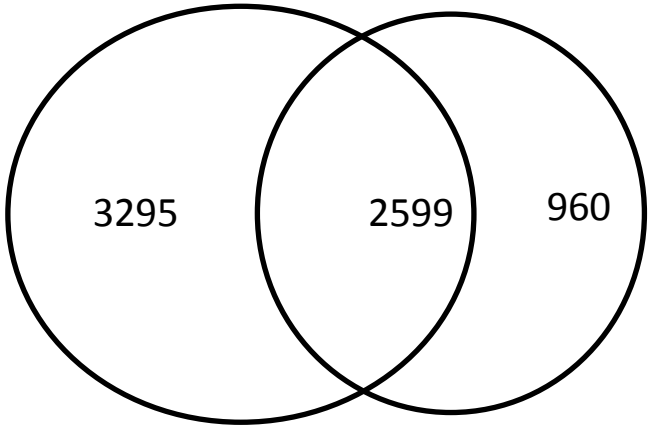
Library size:

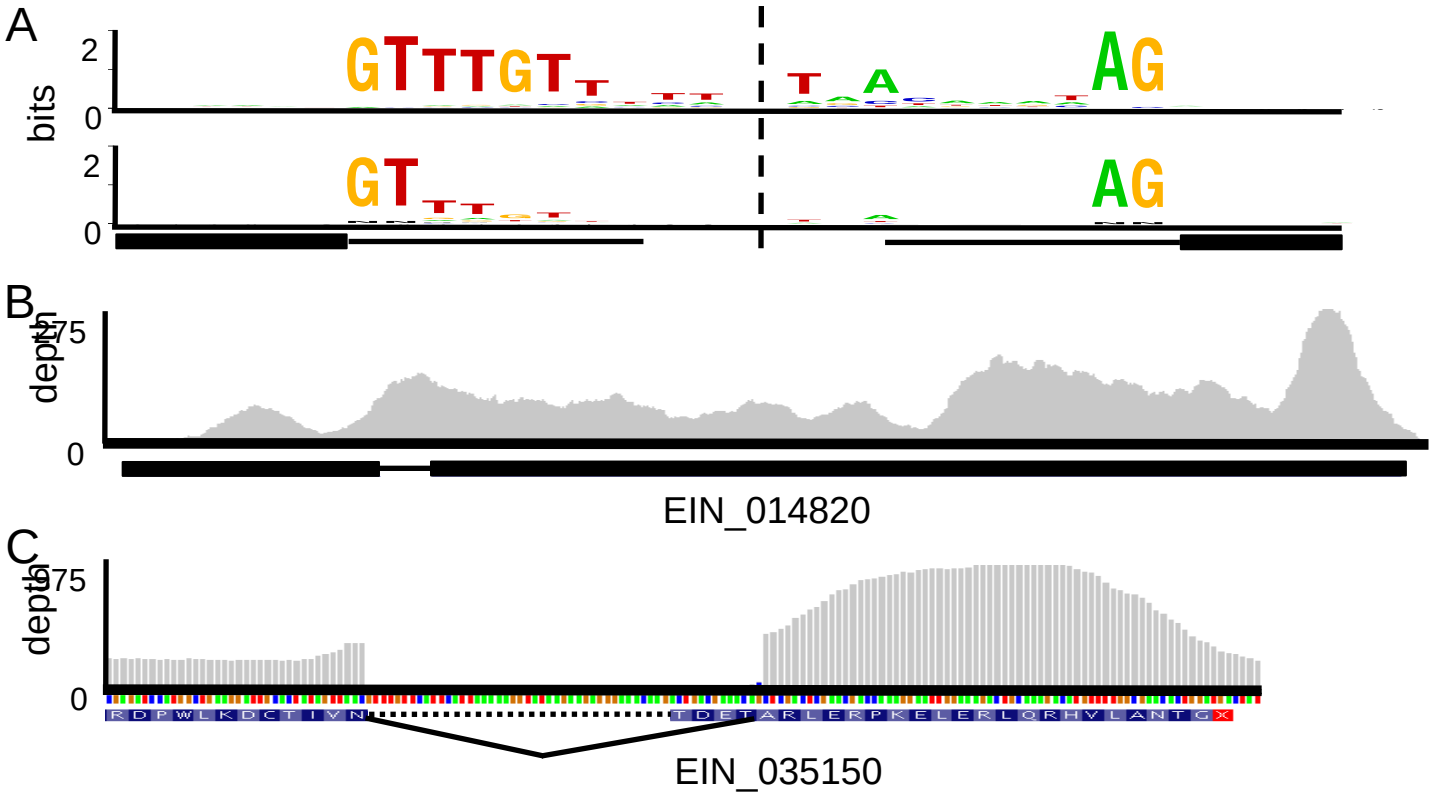
58,449,143	58,660,483	59,468,129	59,024,235	48,908,902	48,106,288/ 41,257,294

2 samples per time point and a third for northern confirmation

Ehrenkaufer et al Genome Biol. 2013 14(7):R77

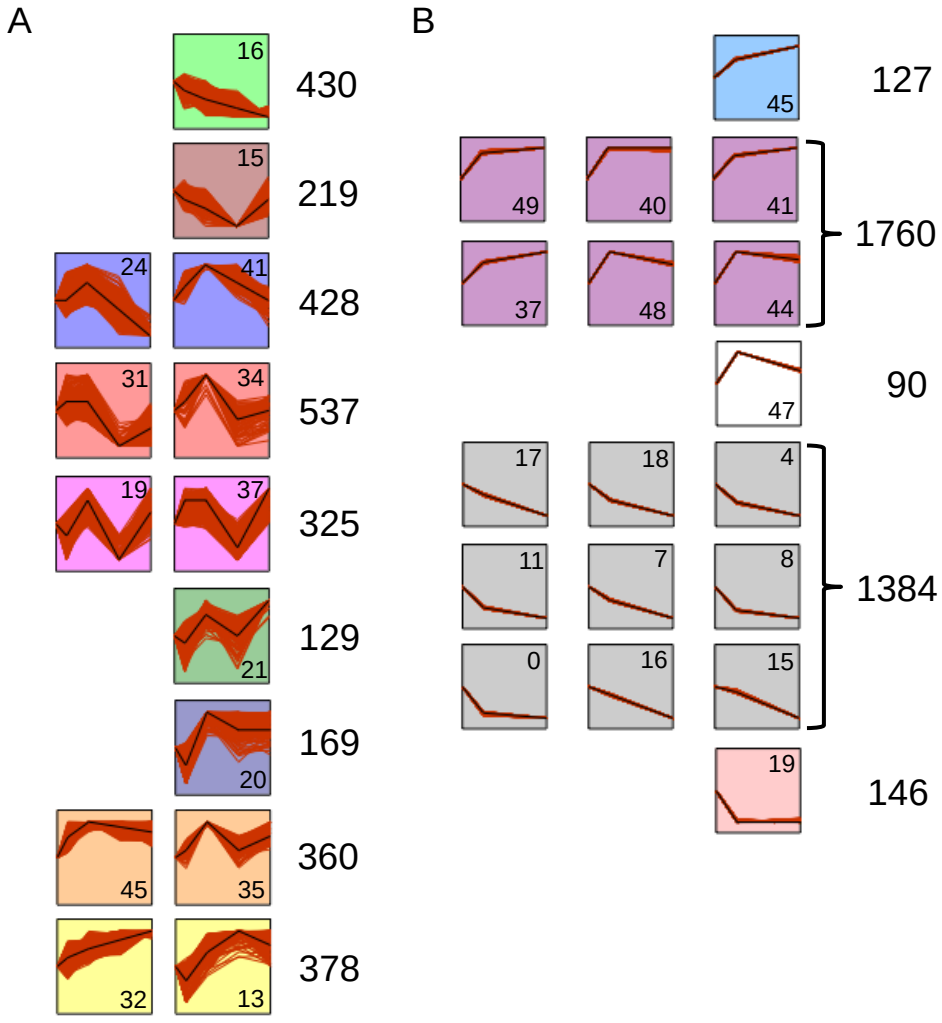
Analysis of *E. invadens* introns





Ehrenkaufer et al Genome Biol. 2013 14(7):R77

Temporal expression profiles



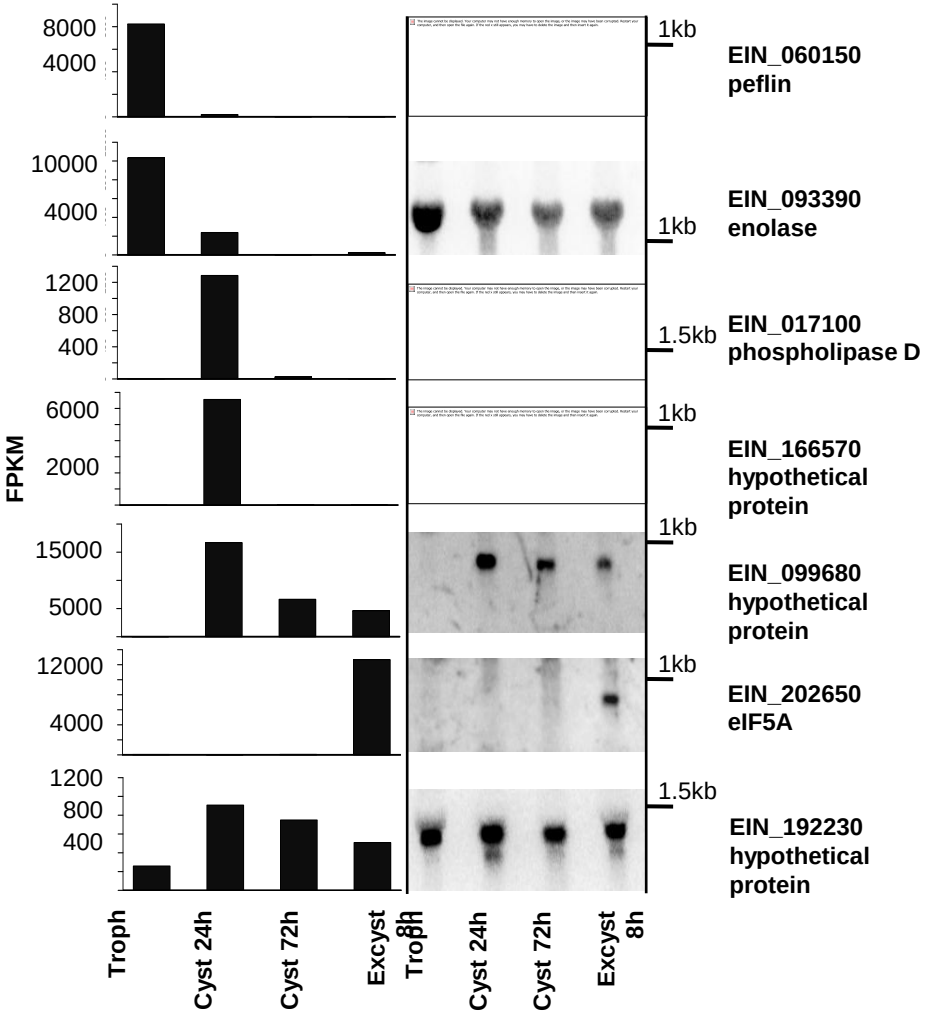
Clustering of gene expression profiles using Short Time-Series Expression Miner (STEM) software

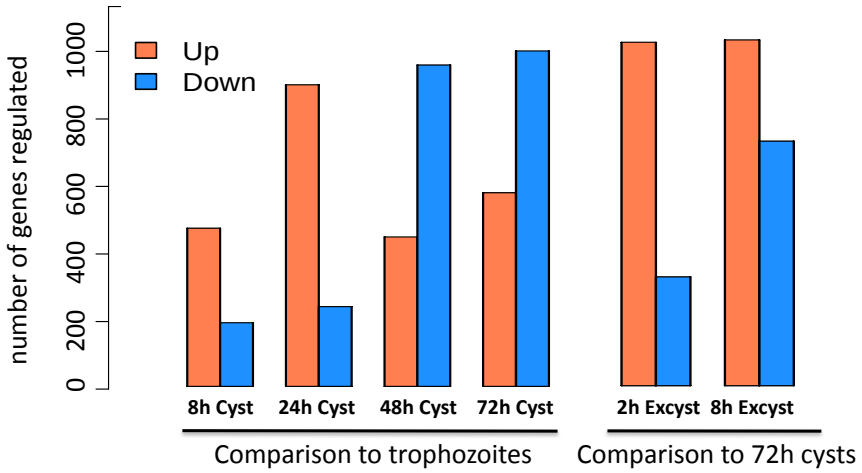
During early encystation we observe a general trend of down regulation

During excystation we observe a general trend of genes being switched on.

Ehrenkaufer et al Genome Biol. 2013 14(7):R77

RNAseq data are confirmed by Northern analysis

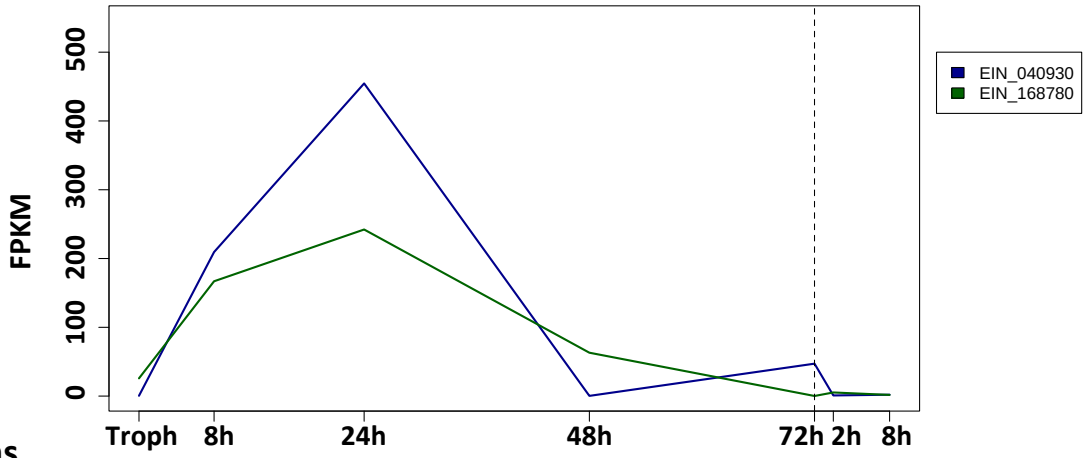




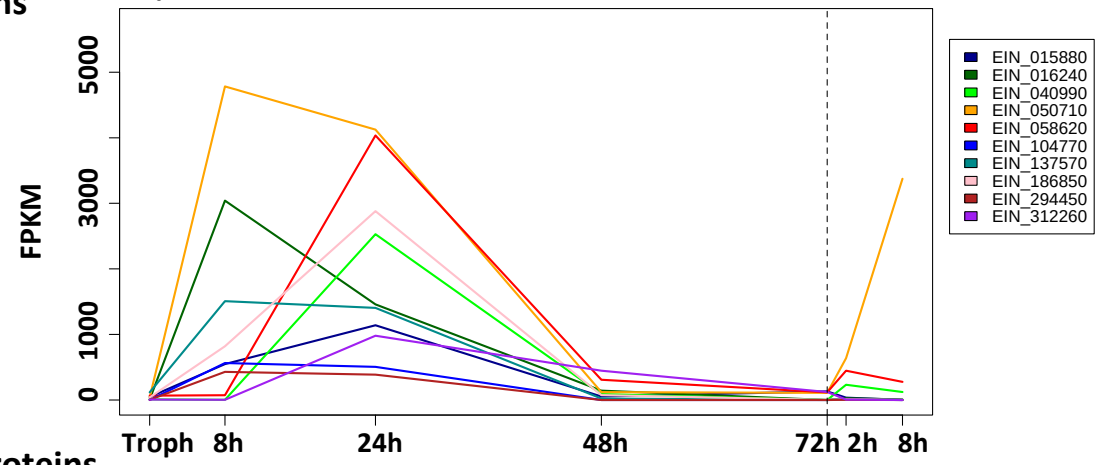
Ehrenkaufer et al Genome Biol. 2013 14(7):R77

Changes in the estimated expression levels of known encystation associated genes

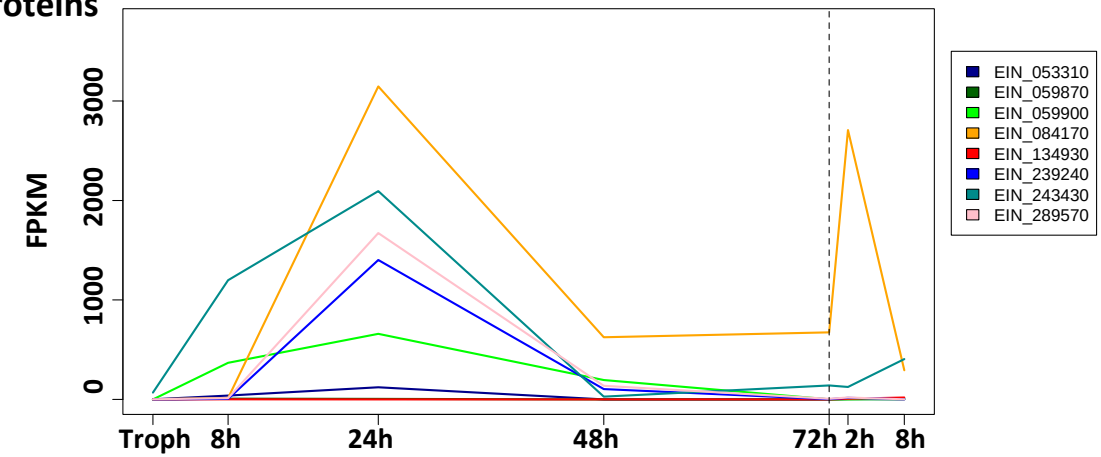
A) chitin synthases



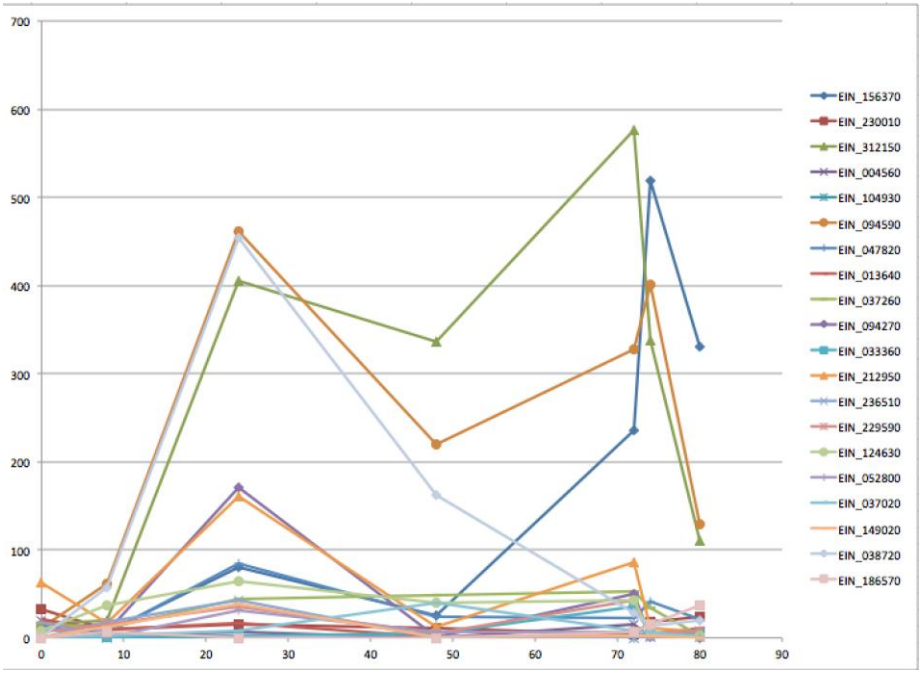
B) chitin binding lectins



C) chitinase domain proteins



expression levels of genes involved in meiosis increase at 24 h after encystation



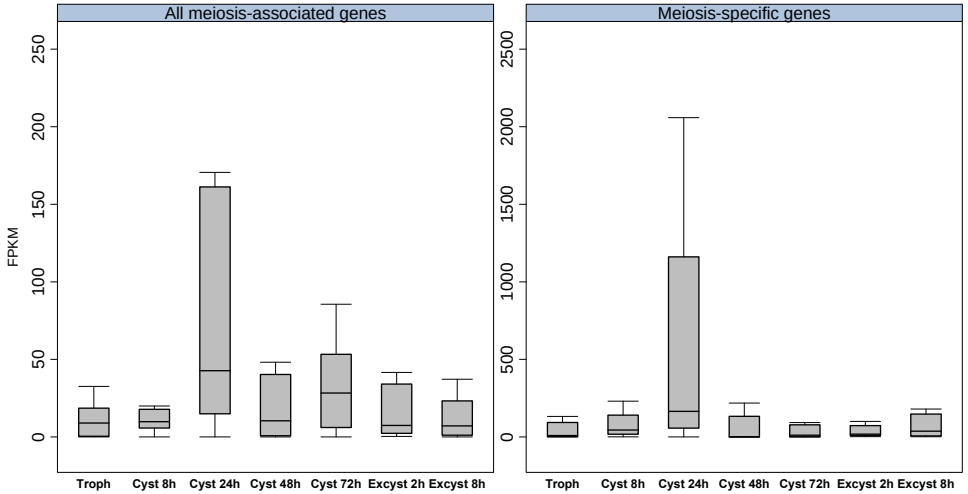


Table 1. Core Meiotic Genes and Some Key Functions of Their Encoded Proteins in Meiosis

Gene	Protein Function(s)
*Spo11**	Transesterase; creates DNA double-strand breaks (DSBs) in meiosis I
Mre11	3′-5′ dsDNA exonuclease and ssDNA endonuclease; forms complex with Rad50 and Xrs2/Nbs1
Rad50	ATPase, DNA binding protein; in a complex with Mre11/Xrs2, holds broken DNA ends together while Mre11 trims
*Hop1**	Synaptonemal complex protein; binds DSBs and oligomerizes during meiotic prophase I
*Hop2**	Forms a complex with Mnd1 to ensure accurate and efficient homology searching during pachytene of meiotic prophase I
*Mnd1**	With Hop2, functions after meiotic DSB formation and is required for stable heteroduplex DNA formation and interhomolog repair
Rad52	Binds DSBs and initiates assembly of meiotic recombination complexes
*Dmc1**	Homolog of strand exchange protein Rad51; promotes interhomolog recombination
Rad51	With Dmc1, catalyzes homologous DNA pairing and strand exchange
*Msh4**	Forms heterodimer with Msh5; interacts with Mlh1/Mlh3; recombination crossover control
*Msh5**	Forms heterodimer with Msh4; interacts with Mlh1/Mlh3; recombination crossover control
Msh2	Forms a heterodimer with Msh3 or Msh6
Msh6	Forms a heterodimer with Msh2; binds base mismatches
Mlh1	Mismatch repair and promotion of meiotic crossing over; interacts with Msh2/Msh6 and Msh4/Msh5; forms heterodimers with Mlh2, Mlh3, and Pms1
Mlh2	Forms a heterodimer with Mlh1; interacts with Msh2/Msh3 and Msh2/Msh6
Mlh3	Forms a heterodimer with Mlh1; interacts with Msh2/Msh3 and Msh2/Msh6 for mismatch repair or with Msh4/Msh5 to promote meiotic crossovers
Pms1	Forms heterodimer with Mlh1 for repair of heteroduplex DNA; interacts with Msh2/Msh3

*denotes meiosis-specific genes. See Supplemental Data for exemplar references for each gene.

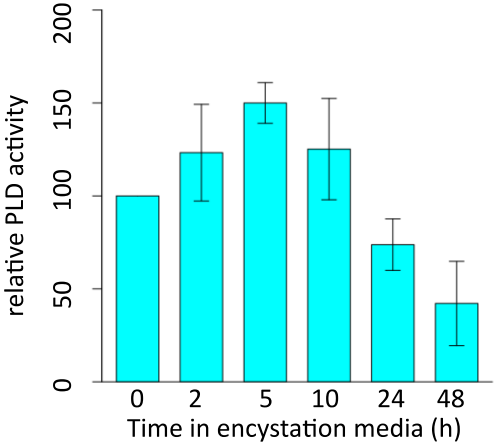
Ramesh et al 2005, Current Biol. 15:185–191

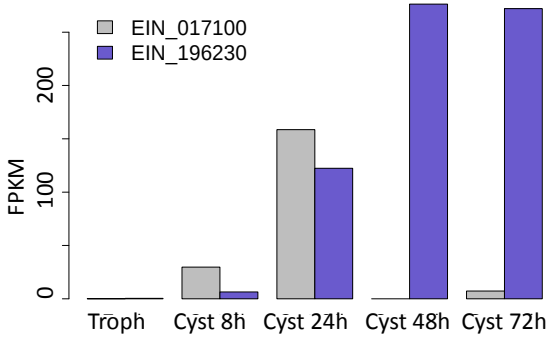
Ehrenkaufer et al Genome Biol. 2013 14(7):R77

Phospholipse D expression and activity increases during encystation

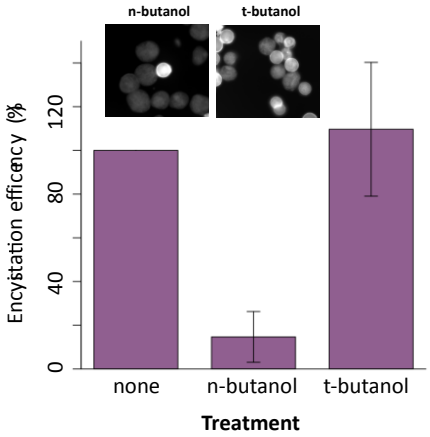
- *E. invadens* has two genes encoding PLDs: EIN_017100 and EIN_196230, both of which are highly upregulated during encystation

- Phospholipase D activity increased during encystation.





- n-butanol was used to inhibit PLD, t-butanol used as a control
- N-butanol significantly reduces encystation in culture.



Ehrenkaufer et al Genome Biol. 2013 14(7):R77

Meiosis and Entamoeba

- Genomic data support the hypothesis that Recombination is occurring in Entamoeba
- We have described the transcriptional changes during encystation and excystation in *E. invadens*.
- We see strong transcriptional evidence for meiosis occurring during the encystation process
- These data further suggests that meiosis is integral to encystation and a common process in *Entamoeba*
- We have identified a lipid signaling molecule (PLD) that is involved in induction of encystation.

Final thoughts

- Genomics is cool
- You don’t need to me a superman/superwoman to be a genome scientist
- Remember you’re a biologist