

Parasites WGS: opportunities and challenges

Simone M. Cacciò

Foodborne and Neglected Parasites Unit

European Union Reference Laboratory for Parasites

Istituto Superiore di Sanità, Rome, Italy

(simone.caccio@iss.it)

Inter-EURLs Working Group on NGS (NEXT GENERATION SEQUENCING)



European Union Reference Laboratory for Legionella pneumophila
<http://eurl-legendaria.anses.it>



June 14-15, ISS, Rome, Italy

The EURL for Parasites

- **Target parasites:**

- **Helminths**

- *Trichinella*
- *Echinococcus*
- *Anisakis*
- *Pseudoterranova*
- *Opisthorchis*
- *Diphyllobotrium*
- *Ascaris*
- *Toxocara*
- and other foodborne helminths

- **Protozoa**

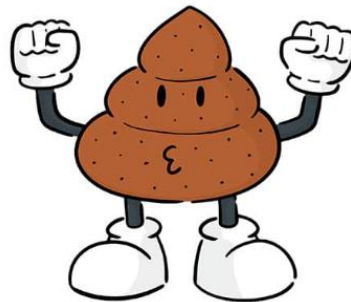
- *Toxoplasma*
- *Giardia*
- *Cryptosporidium*
- *Sarcosystis*
- *Dientamoeba*
- and other foodborne protozoa



Wet-lab: sequencing parasite genomes is not an easy job

- Lack of *in vitro* systems:

what is in the sample is what we have!



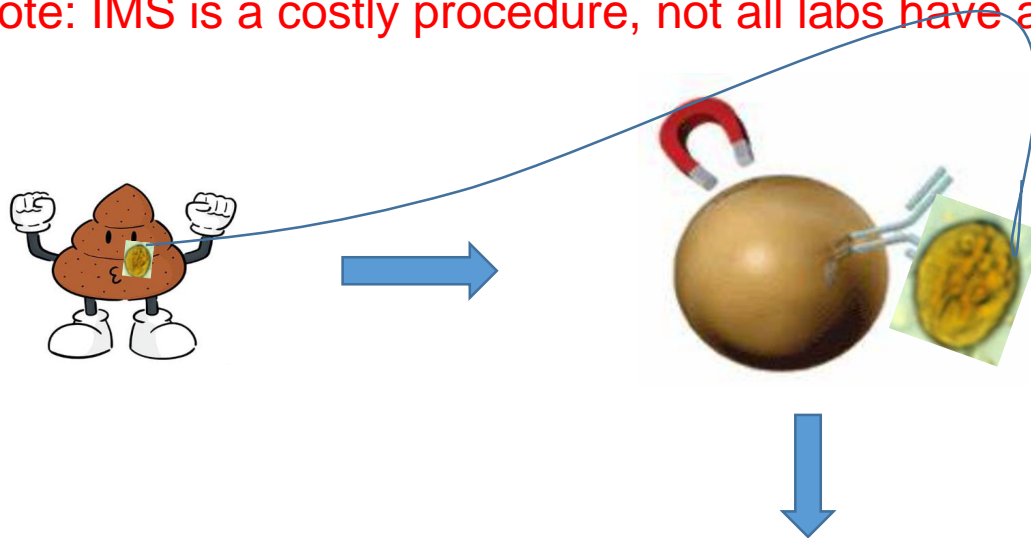
DNA extracted directly from the fecal (or environmental or food) sample cannot be used for WGS, as the fraction of DNA derived from the target organism will be extremely small. It can be used for **metagenomics** studies.

Wet-lab: sequencing parasite genomes is not an easy job

Need highly purified material:

use IMS, FACS or flotation methods on samples
(bias?)

Note: IMS is a costly procedure, not all labs have a FACS



After removal of residual bacteria by bleach treatment, the purity of the recovered cysts is checked and genomic DNA extracted

Note: cysts are very robust and resilient to lysis, so hard methods are needed

Wet-lab: sequencing parasite genomes is not an easy job

Insufficient DNA amount: from parasites purified from clinical samples it is possible to extract nanograms of DNA

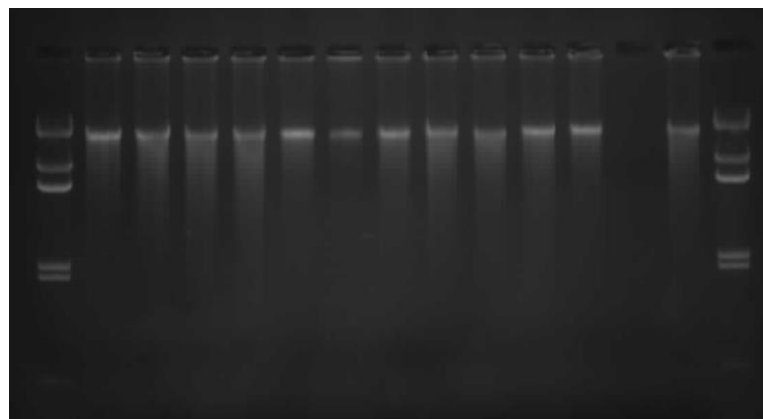
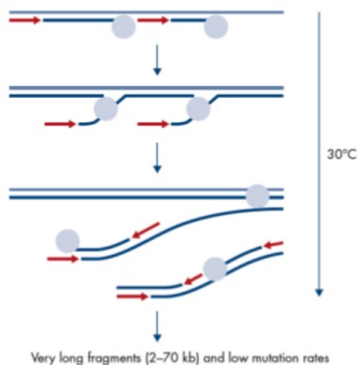
This calls for the use of Whole Genome Amplification (bias?)



Multiple displacement amplification (MDA) by QIAGEN

QIAGEN's REPLI-g technology

- Primers (arrows) anneal to the template
- Primers are extended at 30°C as the polymerase moves along the gDNA or cDNA strand displacing the complementary strand while becoming a template itself for replication
- In contrast to PCR amplification, MDA:
 - Does not require different temperatures
 - Ends in very long fragments with low mutation rates



WGA material is checked again for bacterial contamination (PCR on the 16S ribosomal DNA) and eventually NGS sequenced

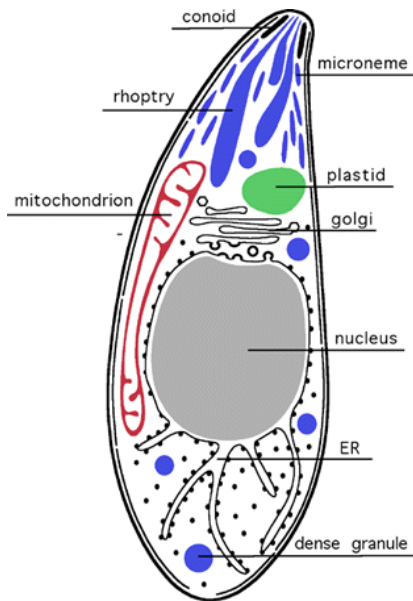
Even if everything goes well.....

Organism	Genome size	N. of chromosomes	Composition GC%	N. of genes
<i>Trypanosoma brucei</i>	26 Mb	>11	46,4	~9000
<i>Trypanosoma cruzi</i>	53 Mb	40	51,7	~19000
<i>Leishmania major</i>	33 Mb	36	59,7	~9400
<i>Giardia duodenalis</i>	12 Mb	5	49,7	~3800
<i>Toxoplasma gondii</i>	66 Mb	12	52,3	~9000
<i>Cryptosporidium parvum</i>	9 Mb	8	30,2	~4000
<i>Plasmodium falciparum</i>	23 Mb	14	19,3	~5600
<i>Schistosoma mansoni</i>	365 Mb	7 + ZW	35,0	~11000
<i>Brugia malayi</i>	94 Mb	4 + XY	27,2	~14000
<i>Onchocerca volvulus</i>	96 Mb	3 + XY	28,3	~12500

Genome size and GC content vary considerably. Worms (helminths) normally have a much larger genome size than protozoa, and are rich in **repetitive** DNA sequences

Let's pick one example: *Cryptosporidium*

- *Cryptosporidium* belongs to the phylum **Apicomplexa**, which comprises human and animal pathogens of great importance (*Plasmodium*, *Toxoplasma*)



All Apicomplexa are unicellular **parasites**, and most are **intracellular** or live in close contact with host cells.

They share the so-called **apical complex**, a sophisticated structure that is fundamental for host cell invasion, which gives the phylum its name.

Why *Cryptosporidium*?

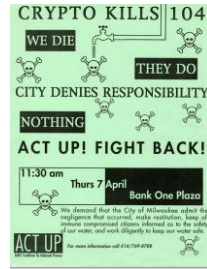
It is a **serious pathogen** for those with an **immature or compromised immune system**

The infection tends to chronicise and disseminate outside of the gut, and is often fatal

It is a pathogen that can infect the general population, particularly via ingestion of **contaminated water and food**. Many outbreaks linked to drinking water, affecting thousands of people

It is the second-most important cause of diarrheal disease among **very young children** living in low-income countries.

There are no effective drugs and no vaccine



Time to tackle
cryptosporidiosis

The little-studied parasite *Cryptosporidium* is a major threat to infants.

Cryptosporidium: essential background

- Many species infect humans
(*C. parvum* and *C. hominis* prevalent)
- Globally distributed
- Complex epidemiology
- Massive waterborne outbreaks
- Immunocompromised at high risk
- Pediatric infection
- Disease burden in developing countries
- Few treatment options and no vaccine
- Lack of (simple) animal models



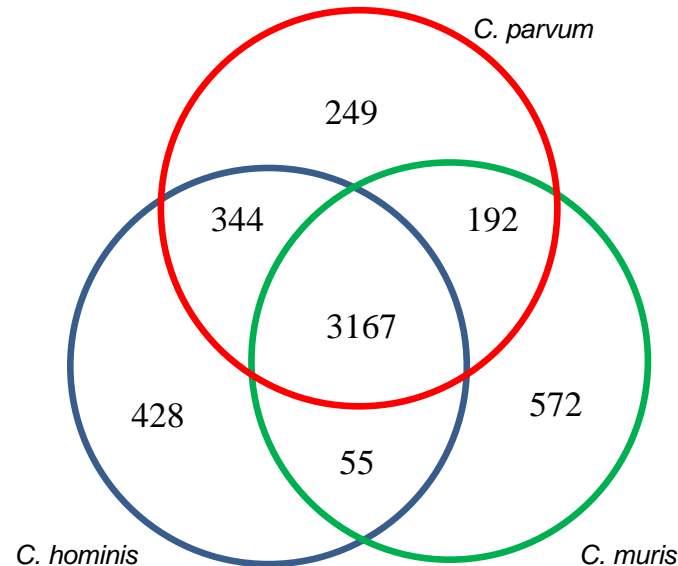
In young children, the parasitic infection cryptosporidiosis is one of four leading causes of severe diarrhea.

Time to tackle
cryptosporidiosis

The little-studied parasite *Cryptosporidium* is a major threat to infants.

Genomics: hard facts

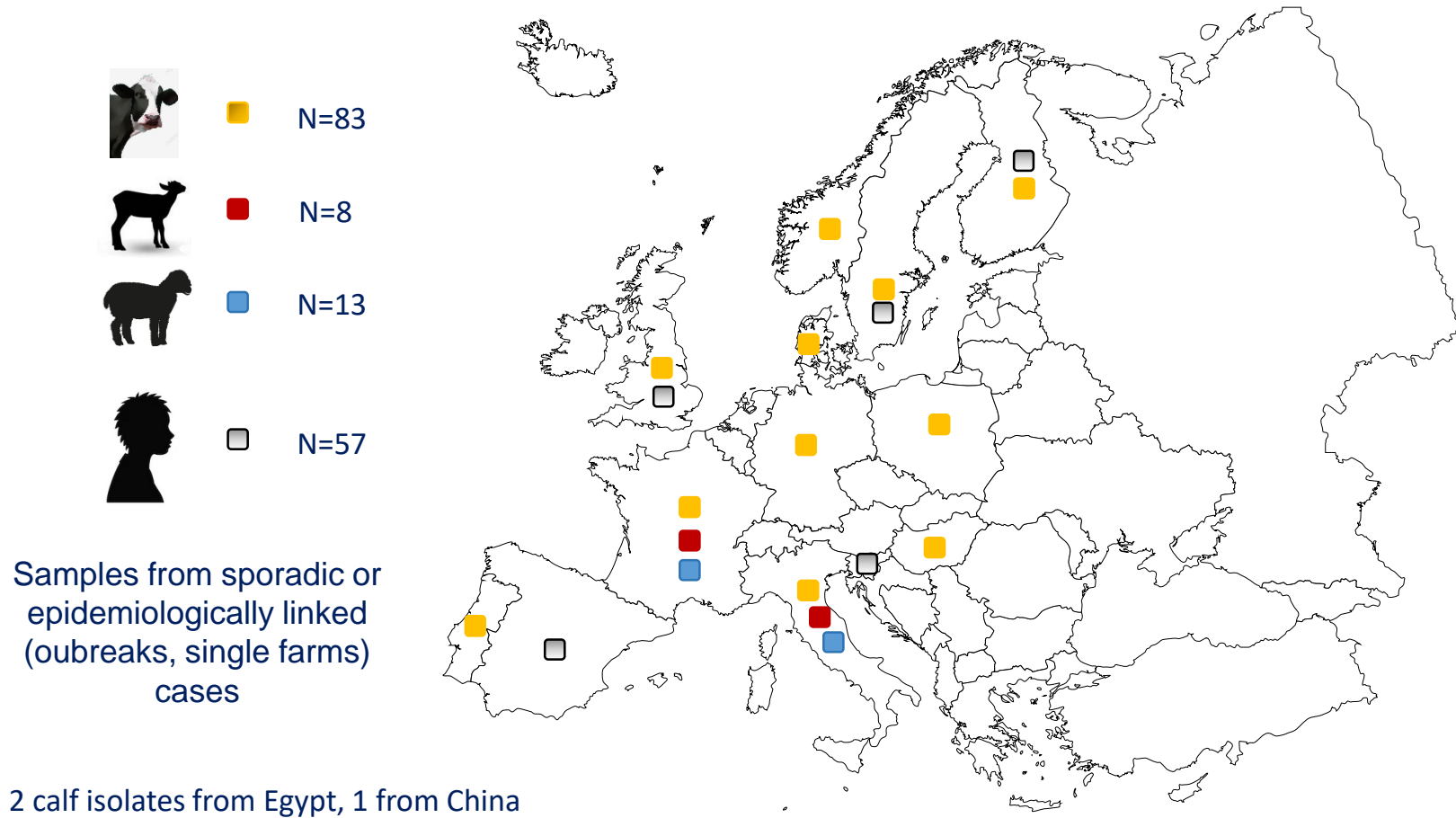
- Small genome (9 Mb)
- Extremely streamlined metabolic pathways
- Organized in 8 linear chromosomes
- 75% annotated as protein-coding (1/3 as hypothetical proteins)
- *C. hominis* and *C. parvum* genomes are largely syntenic



Comparative genomics: what questions?

- How much variation exists at the genome level?
- Population structure (and factors influencing it)
- How similar are genomes from outbreaks?
- It is possible to infer transmission routes?
- How much recombination?
- Virulence factors, genes under selection
- Search for highly polymorphic regions to improve classical genotyping

The dataset of this study: a collection of European isolates of *Cryptosporidium parvum*



How to process WGS data

(Illumina 2x150 bp paired reads)

- Quality check and trimming? **YES**
- Assembly? **YES but....**
- Mapping? **YES**
- Search of genetic features on contigs? **YES, but..**
- gene-by-gene VS SNPs? **SNPs matter more**

- **Phylogeny and cluster analyses**
- **Recombination**
- **Genes under selection**

The first challenge

- The wet-lab procedure I summarized should have clarified that we are sequencing a **population of parasites purified from a biological sample**.
- How good was the purification? Are the sequence data **contaminated**? To what extent?
- This can be done by systematic BLAST of ***de novo* assembled contigs** against non-redundant GenBank database or using tools that assign reads taxonomically (e.g., MetaPhlan, Phyloflash)
- So, assembly is useful!

Contamination very variable, from almost nothing to almost everything!

Isolate	MetaPhlan	Phyloflash	GC
FIN1_S34	-	Thermoactinomyces	42%
FIN3_S36	0,001 Lawsonella	-	31%
FIN4_S37	0,1 clostridium	-	30%
FIN14_S47	1% clostridium	Ashaninka	44%
FIN15_S48	-	-	31%
FIN16_S49	-	-	30%
FIN17_S50	-	-	30%
FIN18_S51	NA	Sphingomonas, Bosea, no crypto	50%
FIN19_S52	NA	NA	41%
FIN20_S53	1% sphingo	Sphingomonas, Psychromas no crypto	49%
FIN21_S54	96% Staf	Sphingomonas, Bosea, Psychromas, Staf, no crypto	44%
FIN22_S55	-	Sphingomonas	31%

Purification is an essential step!

The next step: mapping and variant calling

- We use standard methods, that is to say BWA-MEM v.0.7 to align the reads to a reference genome (*C. parvum* IOWA-ATCC) and the HaplotypeCaller module of GATK v.3.7 (with hard filters).
- After joint genotyping of individual variant call format (VCF) files, we excluded SNPs with alleles < 5X coverage, missingness > 50%, MAF < 0.05 and alternate allele depth < 50%.
- In short, we only use **biallelic sites with good coverage**

The second challenge

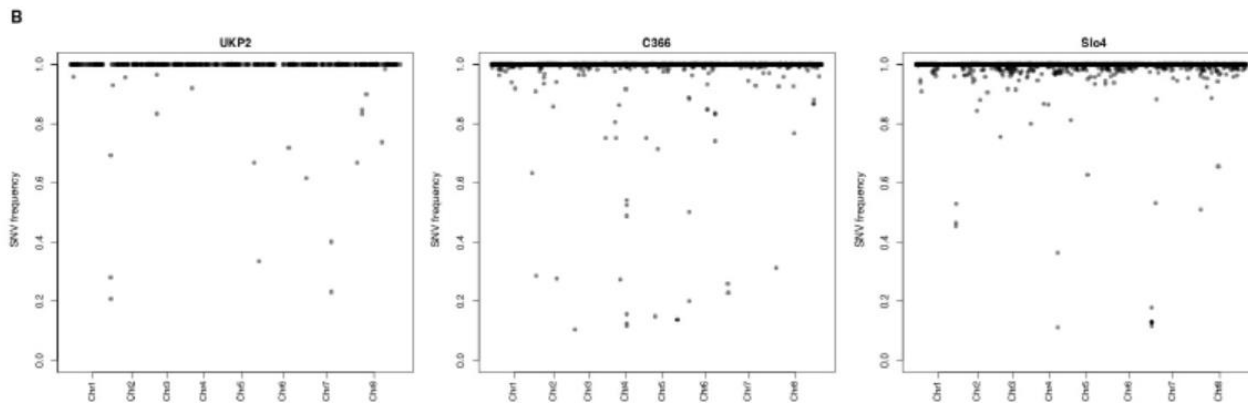
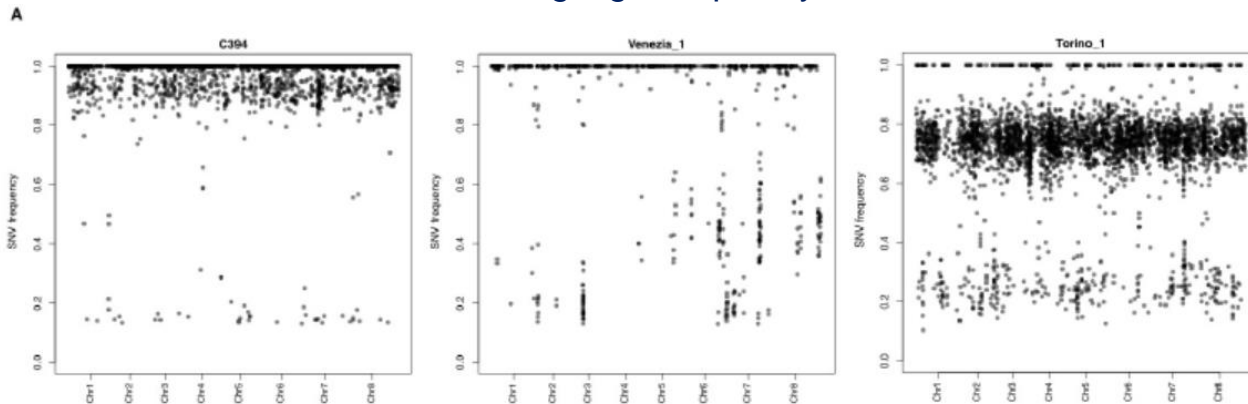
- The wet-lab procedure I summarized should have clarified that we are sequencing a **population of parasites isolated from a biological sample**, and not a clone.
- As **mixed infections** occur in the real world, how can we **estimate the multiplicity of infection**? Remember the cyst is an haploid stage.
- Note: even a single cyst contains 4 organisms that are generated by a meiotic process, thus in principle not necessarily identical...

Multiplicity of infection: the moimix tool

<https://github.com/bahlolab/moimix>

Mixed infections

Three isolates showing high frequency of multi-allelic SNPs

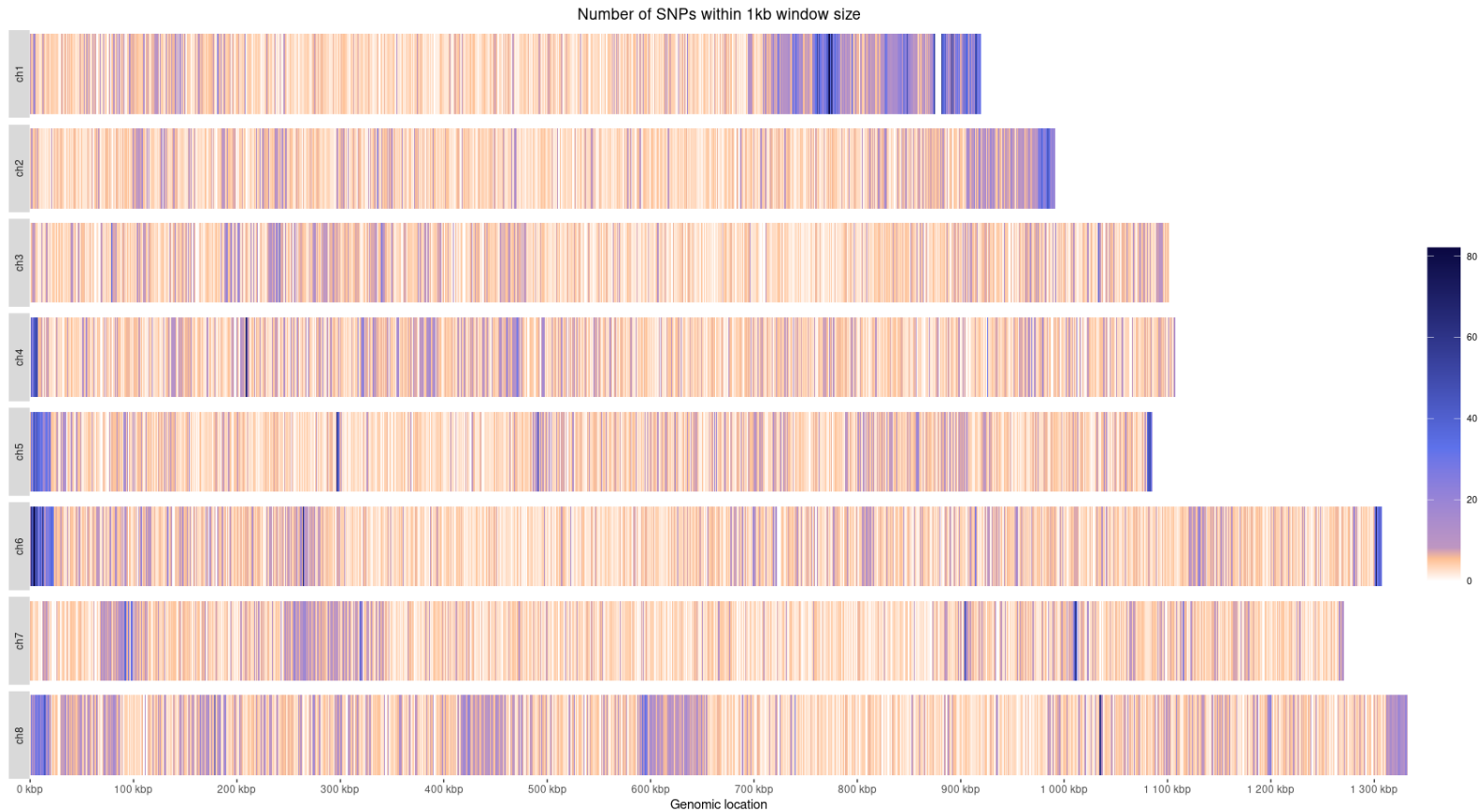


Single infections

Three isolates showing low frequency of multi-allelic SNPs

In pure isolates with haploid genomes, FWS is expected to approach unity. Isolates with $FWS < 0,95$ are excluded, as they likely represent mixed infections

Mapping of reads from «unmixed» genomes to a reference genome

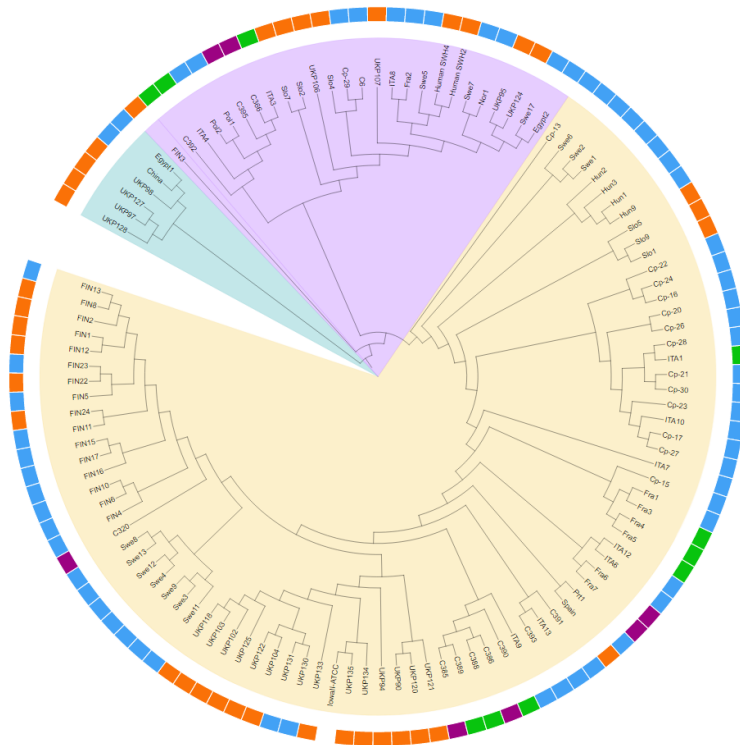
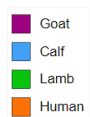


This cumulative plot is based on the **>32,000 SNPs** identified.
Note the **SNP-dense regions close to telomeres**.
The level of genome-wide variability is **modest**

Phylogeny based on genomic SNPs

Three strongly supported clusters

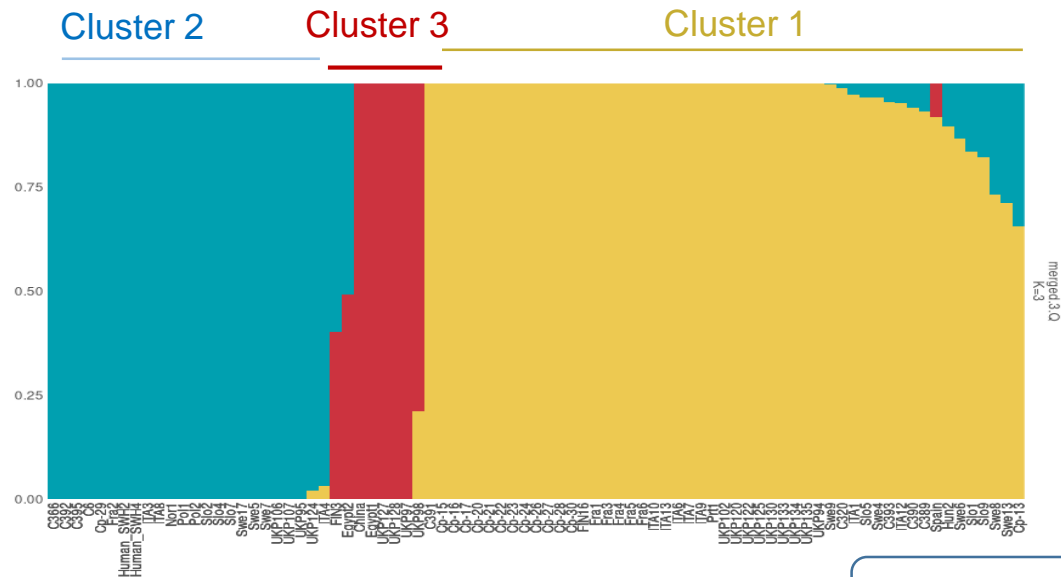
- No cluster by host
- No cluster by geography
- No cluster by previously defined subtypes



Phylogeny inferred by Maximum Likelihood

Additional analyses using genomic SNPs: **STRUCTURE**

The program STRUCTURE implements a model-based clustering method for inferring population structure using genotype data consisting of unlinked markers (SNPs are appropriate). Keep in mind that the model assumes allelic frequency to be at Hardy-Weinberg equilibrium



Signs of admixture among clusters



Recombination?

https://web.stanford.edu/group/pritchardlab/structure_software/release_versions/v2.3.4/html/structure.html

Recombination analyses

- During the **obligatory sexual phase**, recombination between genetically different parasites can occur (reminder: natural mixed infections are not rare)
- To detect recombination, multiple alignments of each chromosome are needed. This is not an easy task, computationally speaking (can be attempted, e.g., using MAUVE)
- One way to go around it is to create «**artificial**» copies of the chromosomes, in which the specific SNPs and InDels (from the VCF) are inserted in the reference chromosome, isolate per isolate.

Recombination analyses

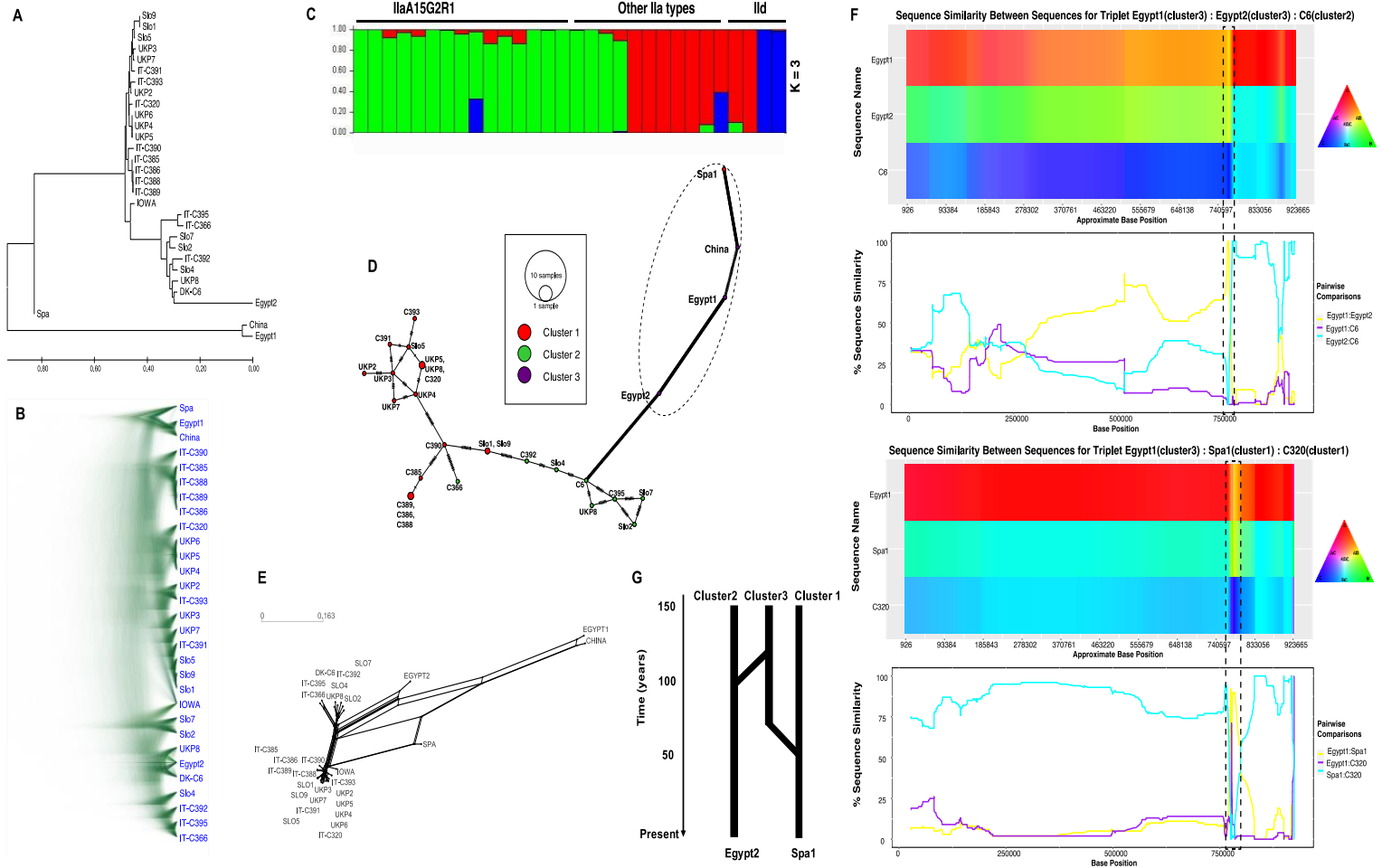
We used the Recombination Detection Program **RDP5** on multiple alignments of each chromosome and identified **190** statistically significant ($p < 0.05$) events. [RDP home page \(uct.ac.za\)](http://uct.ac.za)

Due to the presence of very similar genomes, precise identification of the recombinant and the (minor and major) parents was often difficult.

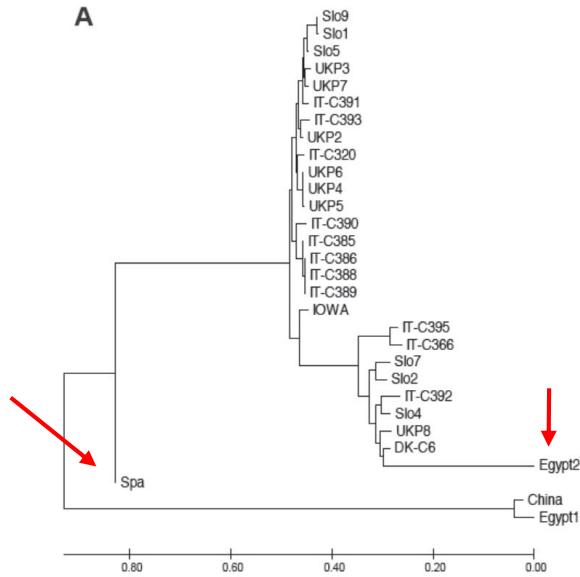
Events were **not randomly distributed**, with chromosomes 1 and 6 showing more events than expected, while chromosome 2 and 7 had less than expected.

Furthermore, regions **close to telomeres** were more often involved in recombination

Insights from single chromosomes: chr. 1

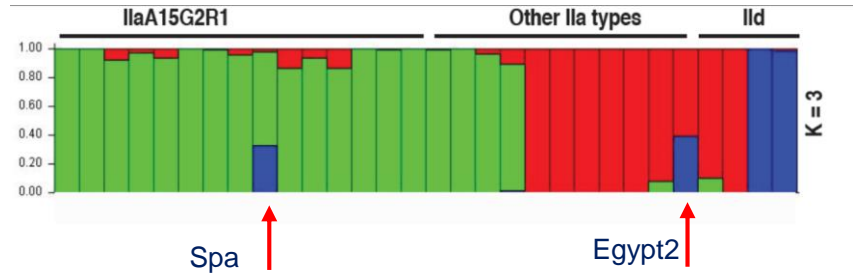


The human isolate Spa and the animal isolate Egypt 2 are hybrids



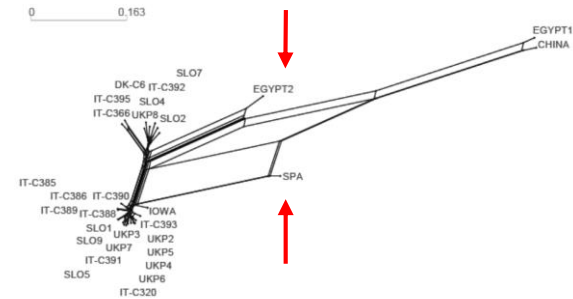
Phylogeny:

Change in the branching of Spa and Egypt2
no longer clustered in their «own» clusters



Structure plot:

Signs of admixture in Spa and Egypt 2



Spit Tree:

Loops connecting Spa and Egypt 2 with cluster 3

Genes under selection

- Genes under **positive or balancing selection** may play important roles in the biology of the pathogen.
- To identify them, nucleotide diversity (π) can be calculated and used to select the top 5% most highly polymorphic genes. These can then be filtered based on Tajima's D values and based on Ka/Ks ratio.
- These calculations can be done using the **PopGenome** package and the **Ka/Ks_Calculator**.
- Additional analyses, e.g., to predict protein localisation or verify enrichment for specific properties (signal peptide, transmembrane domain) can also be performed.
- Finally, one can evaluate whether genes located within recombination breakpoints have specific properties compared to those located outside of these regions.

Conclusions

- Besides the wet-lab challenges, I hope I did show you that there are common steps in the analytical workflow that are applicable to any pathogen.
- And that there are analyses that matter more when studying parasitic pathogens, too.

- No webservers available to process parasite WGS data yet
- Commercial software focus on bacteria/virus, too
- We are working to provide a solution.....stay tuned!

Questions?

Feel free to contact me at
simone.caccio@iss.it



[Microb. Genom.](#) 2021 Jan; 7(1): mgen000493.

PMCID: PMC8115899

Published online 2020 Dec 23. doi: [10.1099/mgen.0.000493](https://doi.org/10.1099/mgen.0.000493)

PMID: [33355530](https://pubmed.ncbi.nlm.nih.gov/33355530/)

Comparative genomics revealed adaptive admixture in *Cryptosporidium hominis* in Africa

Swapnil Tichkule,^{1, 2, 3} Aaron R. Jex,^{1, 4} Cock van Oosterhout,^{5, 7} Anna Rosa Sannella,⁶ Ralf Krumkamp,^{7, 8} Cassandra Aldrich,^{7, 8, 9} Oumou Majga-Ascofare,^{7, 8, 10} Denise Dekker,^{7, 8} Maike Lamshöft,^{7, 8} Joyce Mbwana,¹¹ Njari Rakotozandrainy,¹² Steffen Borrmann,^{13, 14} Thorsten Thys,^{7, 9} Kathrin Schuldt,^{7, 9} Doris Winter,^{7, 8} Peter G. Kreamer,^{13, 14} Kwabena Opong,¹⁰ Prince Manouana,¹³ Mirabeau Mbong,¹³ Samwel Gesase,¹¹ Daniel T. R. Minja,¹¹ Ivo Mueller,^{1, 3} Melanie Bahlo,^{1, 3} Johanna Nader,¹⁵ Jürgen May,^{7, 8} Raphael Rakotozandrainy,¹² Ayola Akim Adegnika,^{13, 14} John P. A. Lusingu,¹¹ John Amuasi,¹⁰ Daniel Eibach,^{7, 8} and Simone Mario Caccio^{6, *}

MOLECULAR ECOLOGY

SPECIAL ISSUE | [Full Access](#)




Recent genetic exchanges and admixture shape the genome and population structure of the zoonotic pathogen *Cryptosporidium parvum*

Giulia I. Corsi, Swapnil Tichkule, Anna Rosa Sannella, Paolo Vatta, Francesco Asnicar, Nicola Segata, Aaron R. Jex, Cock van Oosterhout  Simone M. Caccio  ... [See fewer authors](#) ^



Volume 39, Issue 4
April 2022

Global Population Genomics of Two Subspecies of *Cryptosporidium hominis* during 500 Years of Evolution

Swapnil Tichkule , Simone M. Caccio , Guy Robinson, Rachel M. Chalmers, Ivo Mueller, Samantha J. Emery-Corbin, Daniel Eibach, Kevin M. Tyler, Cock van Oosterhout , Aaron R. Jex 