Complexities of parasite genomes for high-throughput data interpretation: *Leishmania* as an example

Fatma Guerfali, PhD
fatma.guerfali@gmail.com
Institut Pasteur de Tunis
Laboratory of Transmission, Control and Immunobiology of Infections
Group Immunobiology of Infections

One health genomics:

why do pathogenic-related knowledge matter for human health ?

One health genomics: why pathogenic diseases matter for human health

Zoonotic pathogens have serious consequences for human health.

- A single transmission of a virus plausibly from a bat to a boy playing by an old, dead tree has led to the largest outbreak of **Ebola virus** disease ever, claiming over 11,000 lives in West Africa.

- Similarly, the **MERS coronavirus**, seemingly endemic in dromedary camels in the Arabian Peninsula, has repeatedly spilled over to humans, causing numerous outbreaks — the latest : hundreds of patients in South Korea and China after infection of a single traveller.

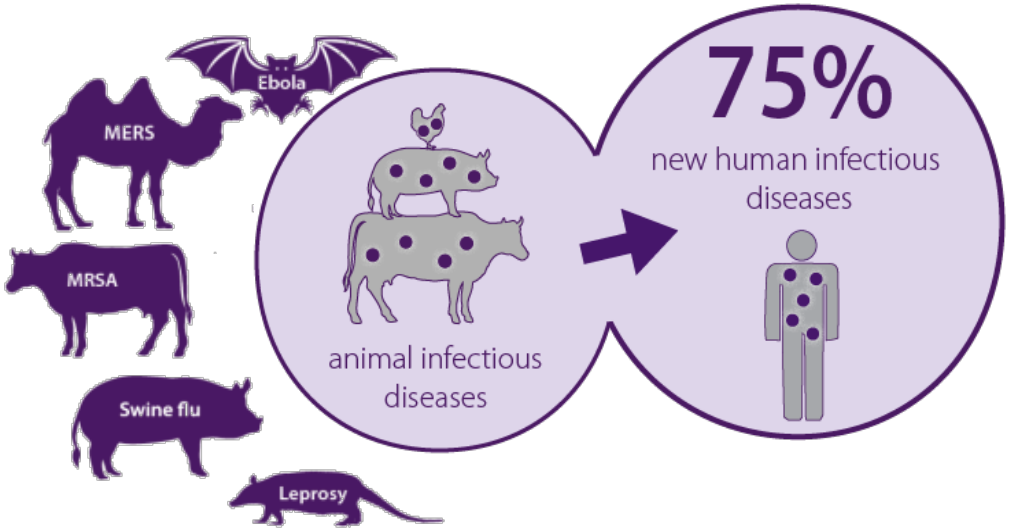
- In 2009, a new lineage of **swine H1N1 influenza** emerged in North America, creating the first pandemic of the 21st century and establishing a new seasonal lineage of influenza A virus.

- Over a million cases of ***Salmonella*** infection occur in the USA each year, and, when outbreaks are large enough to warrant investigation, they are often linked to sources of food production.

Gatdi et al., 2015

One health genomics: why pathogenic diseases matter for human health

- Animals are the source of around **75% of newly emerging human infectious diseases**



- The use of antibiotics to treat bacterial infections in livestock means that if these infections are transmitted to humans they may **already be resistant to many of the antibiotics we use to treat them**

- **Implementing a genomic cross-species surveillance (one health) would enable earlier detection of pathogens and their transmission within and between species**

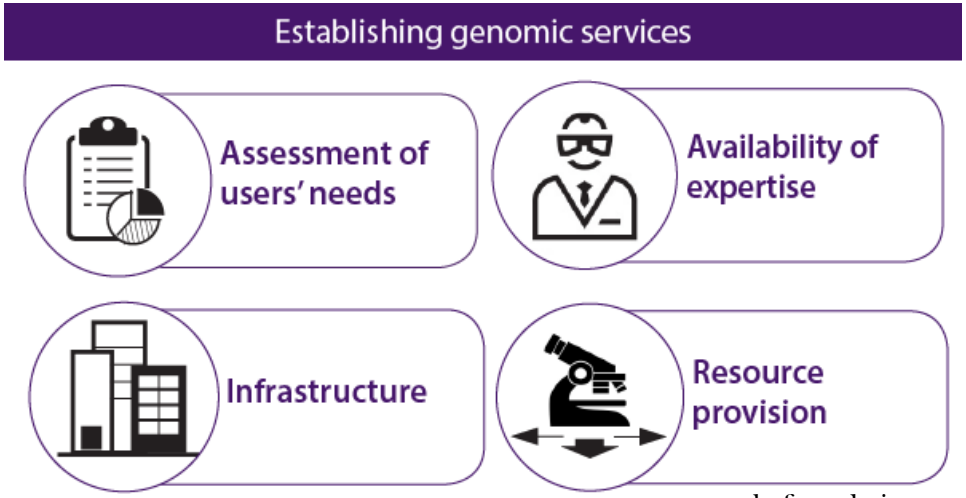
www.phgfoundation.org
Rands, Raza and Luheshi, 2015

One health genomics: why pathogenic diseases matter for human health

- Genomic technology is likely to play a significant role in infection control

Infection control is the principal area where next generation sequencing (NGS) technologies could be used now for infectious disease management and surveillance, complemented by established techniques

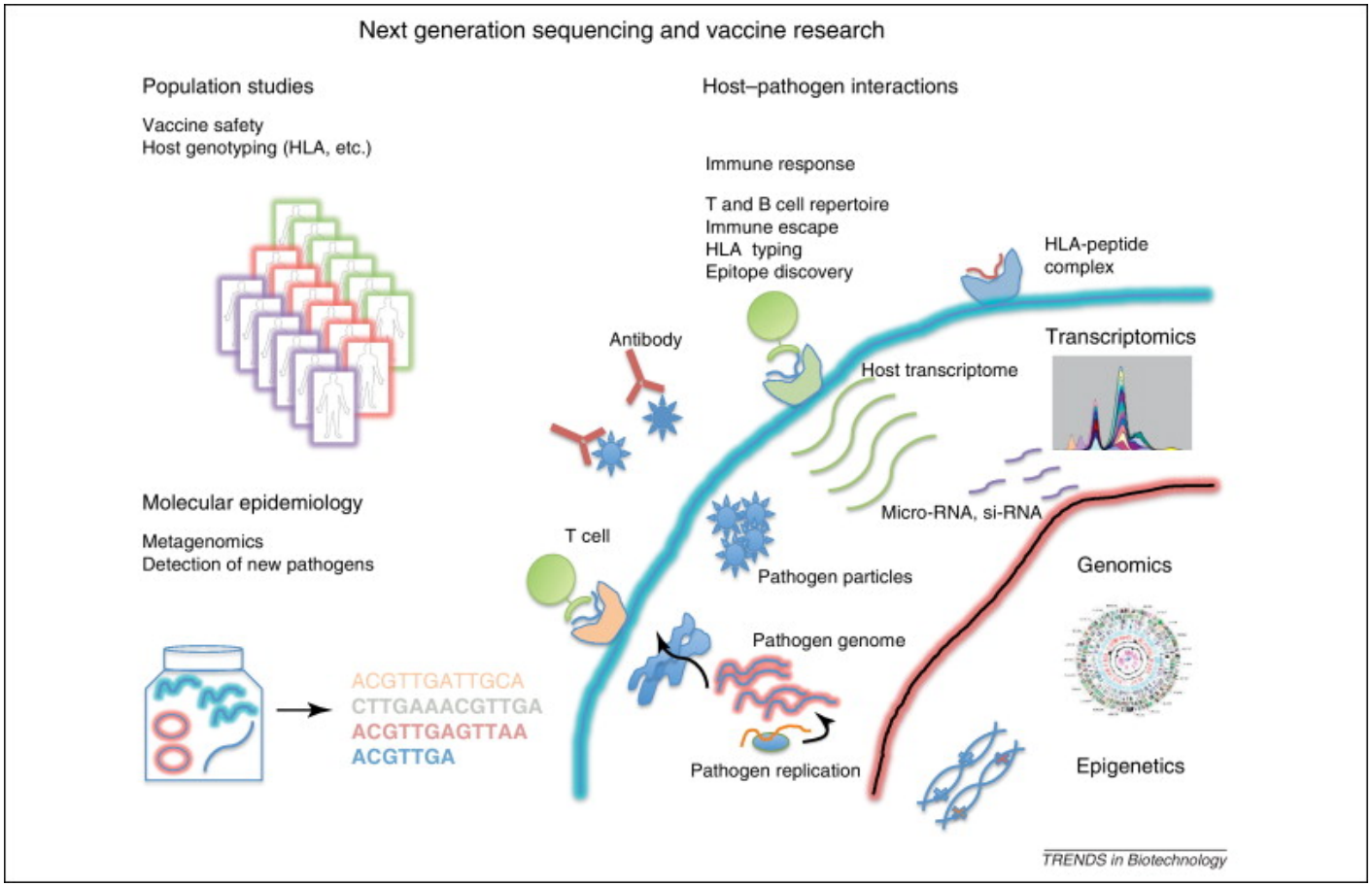
- Hospitals and other health providers will need to make individual decisions about which pathogens they sequence and when sequencing is needed



www.phgfoundation.org
Blackburn, 2015

One health genomics: why pathogenic diseases matter for human health

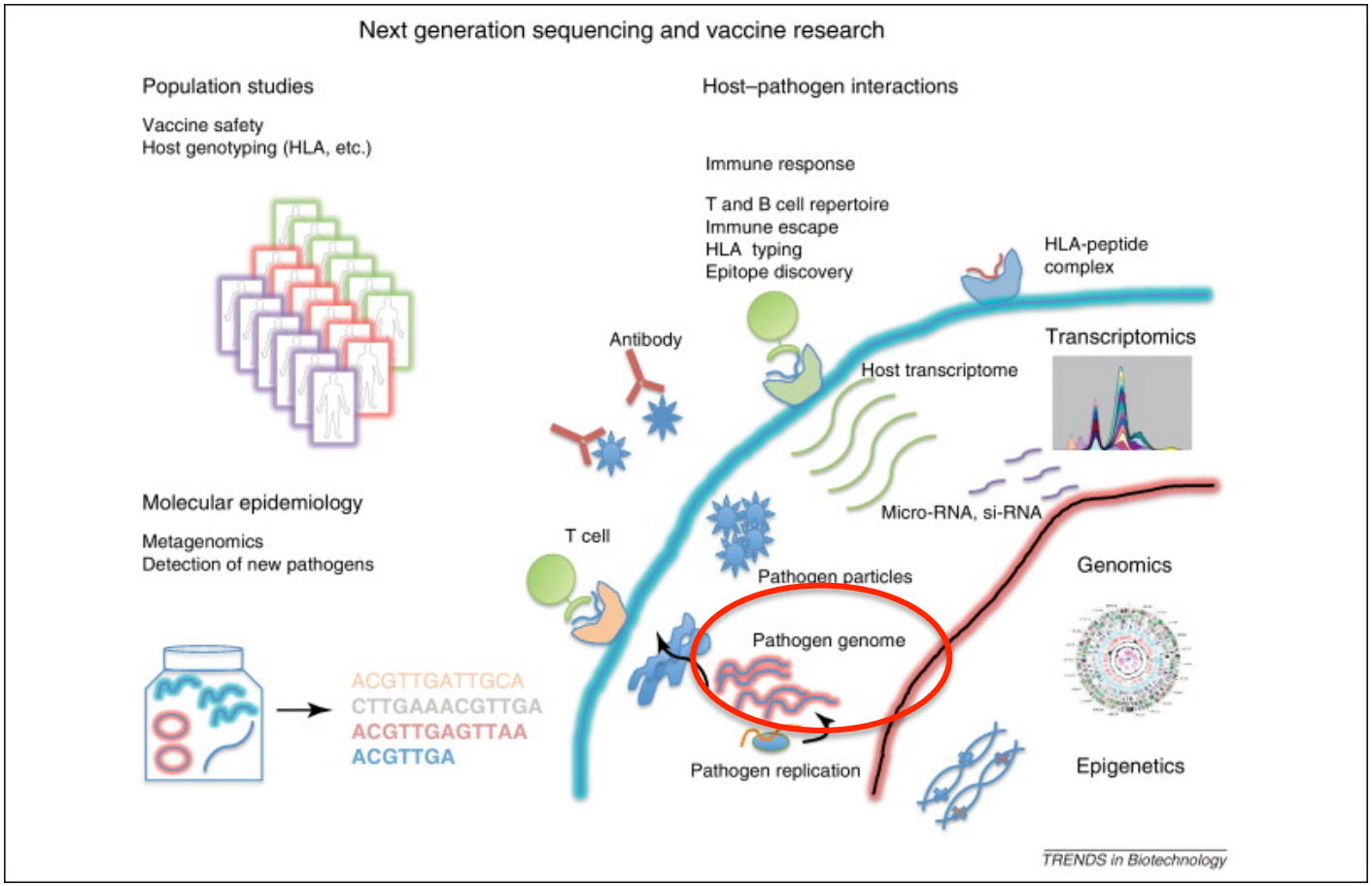
- Sequencing is already being used in many contexts to uncover the source and transmission patterns of outbreaks : e.g. Vaccine research



www.phgfoundation.org
Blackburn, 2015
Luciani et al., 2012

One health genomics: why pathogenic diseases matter for human health

- Pathogen whole genome sequencing (WGS) has several advantages over conventional methods for diagnosing pathogen infections and characterising outbreaks, namely rapid diagnosis, high sensitivity, and flexible analysis



www.phgfoundation.org
Blackburn, 2015
Luciani et al., 2012

How to study genomics ?

- DNA Sequencing (DNA-Seq) is the process of reading the nucleotides present in DNA : **determining the precise order of nucleotides within a DNA molecule.**

- DNA-Seq generally refers to any NGS method or technology that is used to determine the order of the four (ATCG) bases in a strand of DNA.

- There are 2 main types of sequencing technologies that are used today: Sanger sequencing and Next-Generation Sequencing (NGS). Each of these technologies has utility in today's genetic analysis environment.

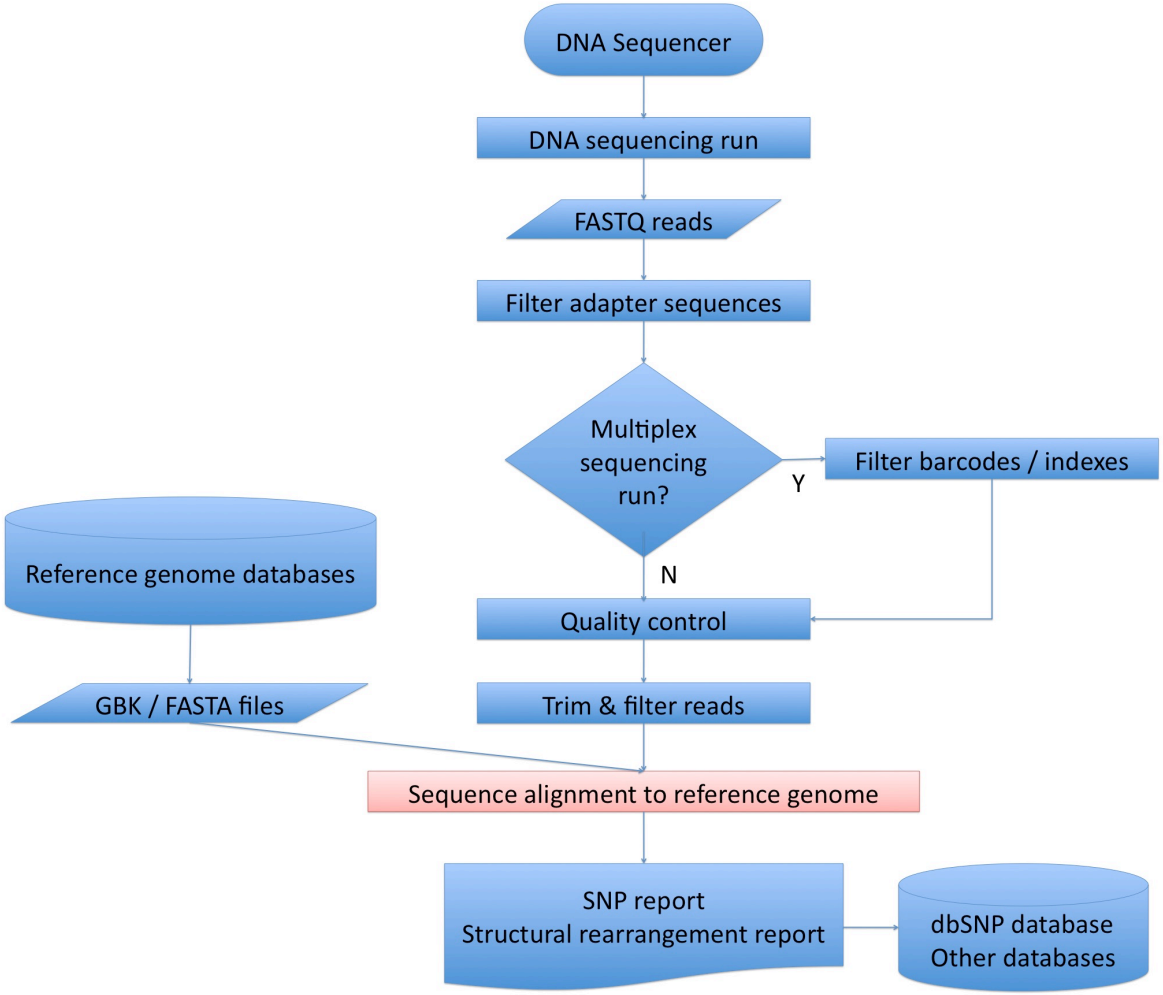
What is DNA-Seq ?

- DNA-Seq is used as an effective sequencing strategy after the advent of rapid DNA sequencing methods that has greatly accelerated biological and medical research and discovery : *de novo*...

- **DNA-Seq may be used to determine the sequence of individual genes, larger genetic regions, full chromosomes, or entire genomes.**

- 'DNA-Seq' and other 'Seq' technologies allow to cover genome complexity : genomic DNASeq, Methyl-Seq, ChIP-Seq, exome sequencing...

DNA-Seq : Pipeline of analysis



The NGS Technologies applications

- Since their development, next-generation sequencing (NGS) technologies have gained increasing attention with a considerable potential application in both diagnostic and public health microbiology (Salipante *et al.*, 2013).

The NGS Technologies applications

Why is it important to have access to *pathogen* genomes?

- As soon as **whole-genome sequencing** entered the scene in the mid-1990s and demonstrated its use in revealing the **entire genetic potential of any given microbial organism**, this technique immediately revolutionized the way **pathogen** (and many other fields of) research was carried out.

- Advances in science and technology aimed at identifying the complete genetic makeup of microorganisms are ushering in a new era for controlling infectious threats.

- *Example:*

*Termed "genomic epidemiology", this approach to infectious disease control was named one of the six **"Areas to Watch in 2012"** by the journal Science on the basis of its transformative potential in **"determining quickly where newly emerging diseases come from, whether microbes are resistant to antibiotics, and how they are moving through a population."***

→ **Applications ?**

The NGS Technologies applications

New ideas and innovative research

In **clinical microbiology**, applications of genome sequencing include
- the development of detection, identification and genotyping tools
- the assessment of antibiotic resistance or virulence repertoires

Recently, thanks to NGS technologies such as the MiSeq (Illumina), Ion Torrent Personal Genome Machine (PGM) (Life Technologies) and 454 GS Junior (Roche) bench-top sequencers, bacterial genome sequencing has become :
- fast (only few hours)
- cheap (only a few hundred US dollars)

→ makes whole-genome sequencing compatible with the routine clinical microbiology workflow.

The NGS Technologies applications

New ideas and innovative research

By using genetic sequencing to examine infectious pathogens, these technologies has/ will revolutionize(d) our ability to:

- diagnose infectious diseases → identify bacterial pathogens
- investigate and control outbreaks → detect laboratory cross-contamination (Köser *et al.*, 2012).
- Identify virulence factors → understand transmission patterns
- develop and target vaccines
- Determine antimicrobial resistance → predict antibiotic resistance
- track the spread of emerging bacteria through epidemiological typing (Price et al., 2013; Sentausa & Fournier, 2013)
- Identify single genomic changes between two isolates (Bertelli & Greub, 2013)
- Comparative genomics

—all with increased timeliness and accuracy and decreased costs!

(http://stacks.cdc.gov/)
(Adapted from web sources)

The NGS Technologies applications

Pathogen Genomics

New ideas and innovative research

Genomic sequence information from cultivated microorganisms is widely used for epidemiological studies → makes whole-genome sequencing compatible with the routine clinical microbiology workflow

Table 1 | **Examples of infectious disease outbreaks that were investigated using next-generation sequencing**

Microorganism	Location	Year	Reference
Carbapenem-resistant *Klebsiella pneumoniae*	USA	2011	112
Clostridium difficile	Worldwide	2013	113
Escherichia coli O104:H4	Germany	2011	114,115
Legionella pneumophila serogroup 1	United Kingdom	2013	116
Methicillin-resistant *Staphylococcus aureus* (MRSA)	United Kingdom	2009	117
Mycobacterium tuberculosis	Canada	2006–2008	118
Vibrio cholerae O1 biovar El Tor	Haiti	2010–2011	119
Arenavirus	Australia	2008	120
Bas-Congo virus	Democratic Republic of the Congo	2009	121
Influenza A virus H1N1	Worldwide	2009	122

Fournier et al., 2013

The NGS Technologies applications

Why is it important to have access to *pathogen* genomes?

NGS could help develop new approaches to reducing Healthcare-Associated Infections and Detecting Antibiotic Resistance

Healthcare-associated infections (HAIs) related deaths occur everywhere in the world (>1million/year in USA ≈ approximately 1 in 20 patients), despite efforts to reduce them:

- Tremendous impact on human health
- Billions of dollars are associated to healthcare costs each year
- New resistant strains are being identified across a spectrum of healthcare-associated infections !!

  methicillin-resistant *Staphylococcus aureus* (MRSA), *Clostridium difficile*, and even more difficult-to-treat, gram-negative bacterial infections such as *Escherichia coli* and *Klebsiella pneumoniae*. These gram-negative infections are increasingly resistant to most available antibiotics and can also pass along genetic materials that enable other bacteria to become drug-resistant.

→ Advances in genomic sequencing can play a critical role in rapidly identifying these infections, tracking their spread, and improving control measures.

(http://www.cdc.gov)

The NGS Technologies applications

Why is it important to have access to *pathogen* genomes?

Examples of the impact of NGS for microbial identification in clinical and public health settings

"Pan-Microbial Diagnosis & Discovery Using Next-Generation Sequencing" (2013)

- Development and validation of a real-time NGS assay for pan-microbial diagnosis ="open," primer-independent technology that enables discovery and reconstruction of novel infectious agents. Because all microbes except prions contain DNA and/or RNA, the full spectrum of disease-causing viruses, bacteria, fungi, and parasites could be readily identified in a single assay.

- Sensitivity to 10-100 viral copies per mL with less than 24 hour turnaround from sample to answer

- Rapid data analysis with the use of cloud-compatible bioinformatics pipelines

The NGS Technologies applications

Why is it important to have access to *pathogen* genomes?

Examples of the impact of NGS for microbial identification in clinical and public health settings

Long et al., 2013:
carried out a one-day exercise to test the feasibility of integrating whole-genome sequencing into the routine workflow of a clinical laboratory.

- They sequenced all isolates received in that day
- Result:
  1. whole-genome sequencing is able to identify all 'mixed-sample' organisms taken from primary isolation plates.
  2. strains with few numbers of reads were easily identified, despite a very low depth of coverage

- **Conclusion:**

identifies fastidious slow growing organisms + rapidly characterizes organisms in mixed cultures → Confirms NGS power to quickly resolve complex bacterial populations

The NGS Technologies applications

Why is it important to have access to *pathogen* genomes?

Examples of the impact of NGS for microbial identification in clinical and public health settings

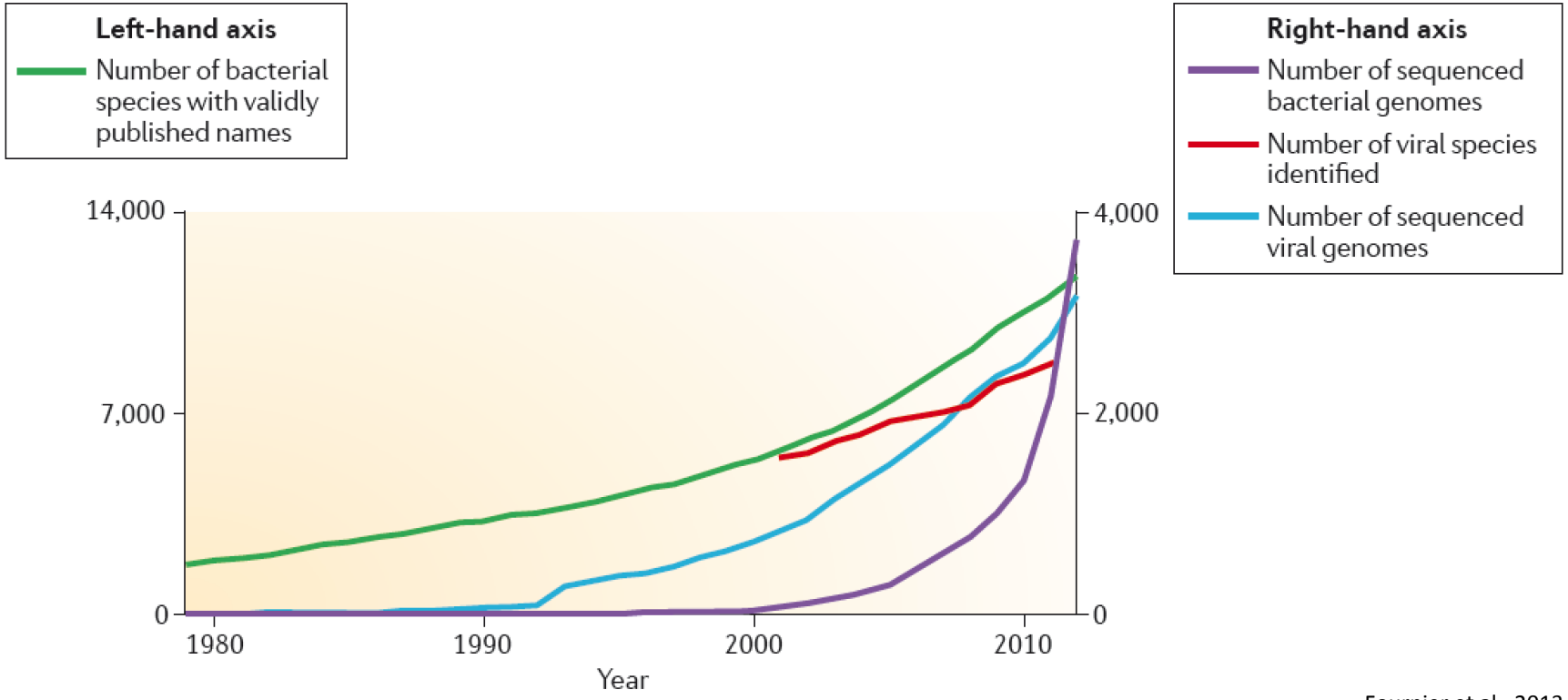
- HistoGenetics to Incorporate MiSeq into HLA Typing Workflow
- "Pan-Microbial Diagnosis & Discovery Using Next-Generation Sequencing" (2013) with open and primer-independent NGS
- PathoQuest to Launch Clinical Trials of NGS-Based Infectious Disease Diagnostic Assay
- MicrobeNet: Creating a system of web-accessible, searchable databases containing detailed, reference information to characterize infectious pathogens. Ensuring near real-time analysis and feedback of information on submitted pathogens for CDC and its local, state, national, and international laboratory partners

http://www.cdc.gov

The NGS Technologies applications

Why is it important to have access to *pathogen* genomes?

The number of identified microbial species from 1979 to 2012. The development of new technologies has had a substantial impact on the number of microbial species that are identified each year.



Fournier et al., 2013

The NGS Technologies applications

Provide adequate references genomes first !!

In order to **provide adequate references for sequencing targets** and for better understanding community composition in metagenomes, a number of efforts have begun to **sample and sequence references** such as:

- the Genomic Encyclopedia of Bacteria and Archaea (Wu et al., 2009)
- the human microbiome effort (Nelson et al., 2010)
- TritrypDB…

Hu et al., 2011

The NGS Technologies applications

Reference genomes

Careful: providing reference genomes might be challenging for pathogens

Knowledge of genome caracteristics and contents for pathogens needed

- GC or AT rich
- Related species identified or not
- Strain-to-strain variability
- Availabilty of a reference genome (comparing pathogens to less, or non-pathogenic nearest neighbors)
- For most genomes, NGS-based draft sequencing is relatively inexpensive and easy compared with the expense and difficulty of complete genome closure.

(Hu et al., 2011)
(Chain et al., 2009)

The NGS Technologies applications

Clinical microbiology

Requirements before introducing it into routine and diagnostic laboratories

Routine analysis of NGS data requires the ***effort and collaboration of specialists*** *from different disciplines*

- Clinicians
- Biologists
- Bioinformaticians
- Programmers
- Statisticians

which is not always feasible in every institution!

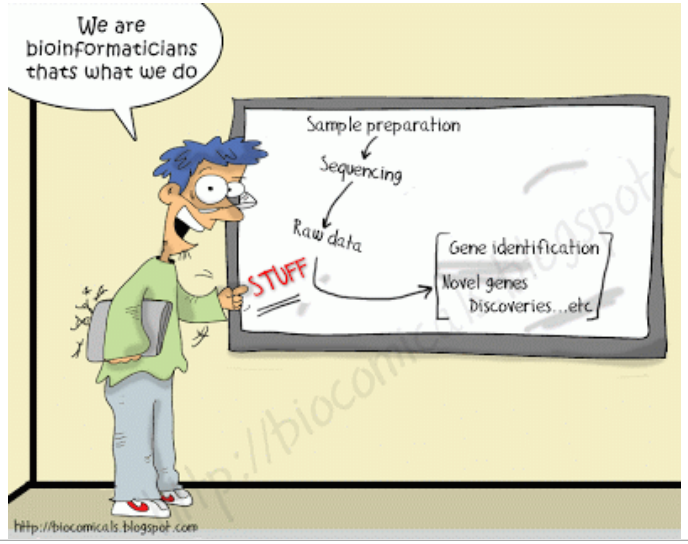
First things first...who is doing this

- Lack of understanding and communication !!!

Biological knowledge

Computational knowledge

Computational solutions



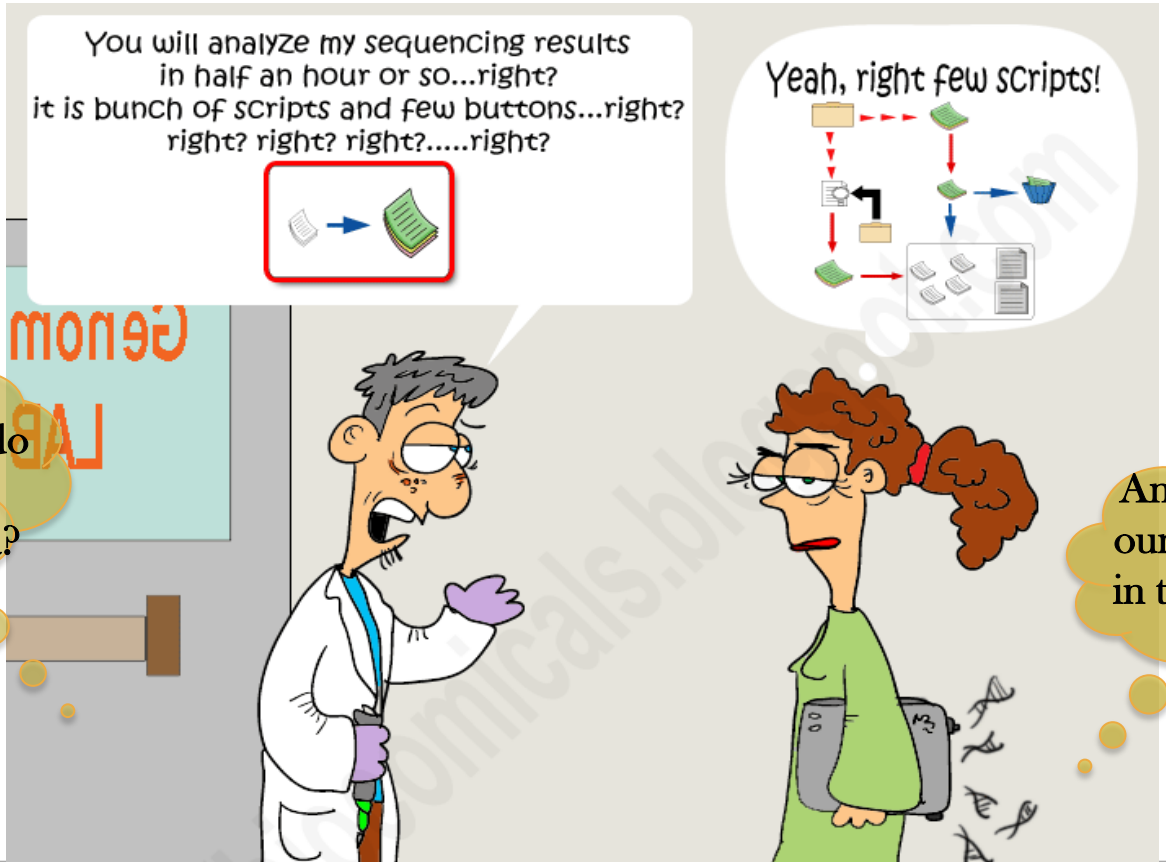
Biological questions

First things first...who is doing this

- Lack of understanding and communication !!!

What Biologists think Bioinformaticians are doing

What Bioinformaticians think Biologists are doing



The NGS Technologies applications

CONCLUSION: Why is it important to have access to *pathogen* genomes?

(Price et al., 2013): Such rapid identification of isolates would:

- help clinicians to initiate a quick and effective therapy
- allows the prompt implementation of targeted infection control practices

However, as for antibiotic resistance:

- NGS technologies will not *completely* replace the phenotypic testing
- would complement it when it comes to investigating abnormal results or detecting resistance with known mechanisms (Köser *et al.*, 2012).

Leishmania genomics:

A human protozoan parasite as an example

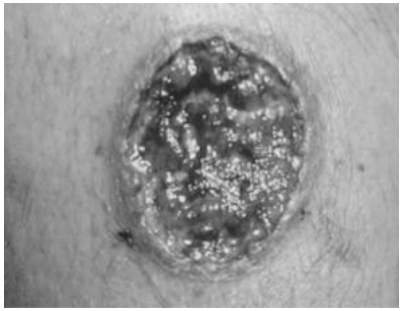
Leishmaniasis : a parasitic disease as an example

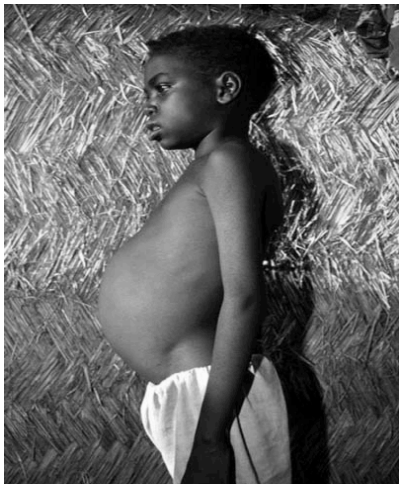
Leishmania are small protozoan parasites.

Their parasitic life cycle includes the sandfly and an appropriate host. Humans are one of these hosts.

Leishmania infection affects :

- Skin → called cutaneous leishmaniasis (CL)
- Mucosal membranes → wide range of appearance, most frequently ulcers. It may cause skin lesions that resemble those of other diseases (leprosy...)
- Viscera → fever, weight loss, enlargement of spleen and liver, abnormal blood test (anemia, leukopenia, thrombocytopenia)

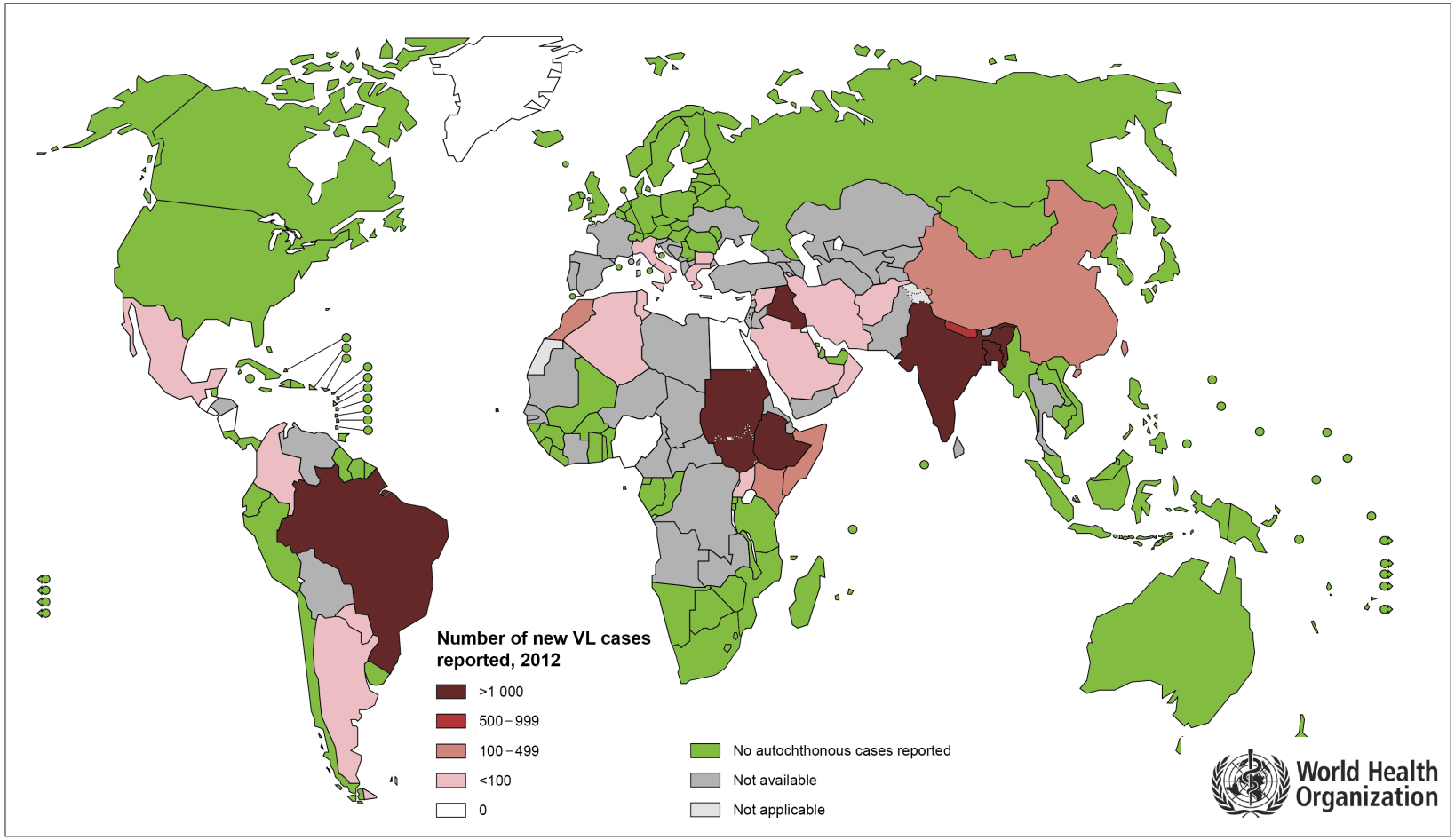




Adapted from Georgina Lewis; http://slideplayer.com/slide/5808585/

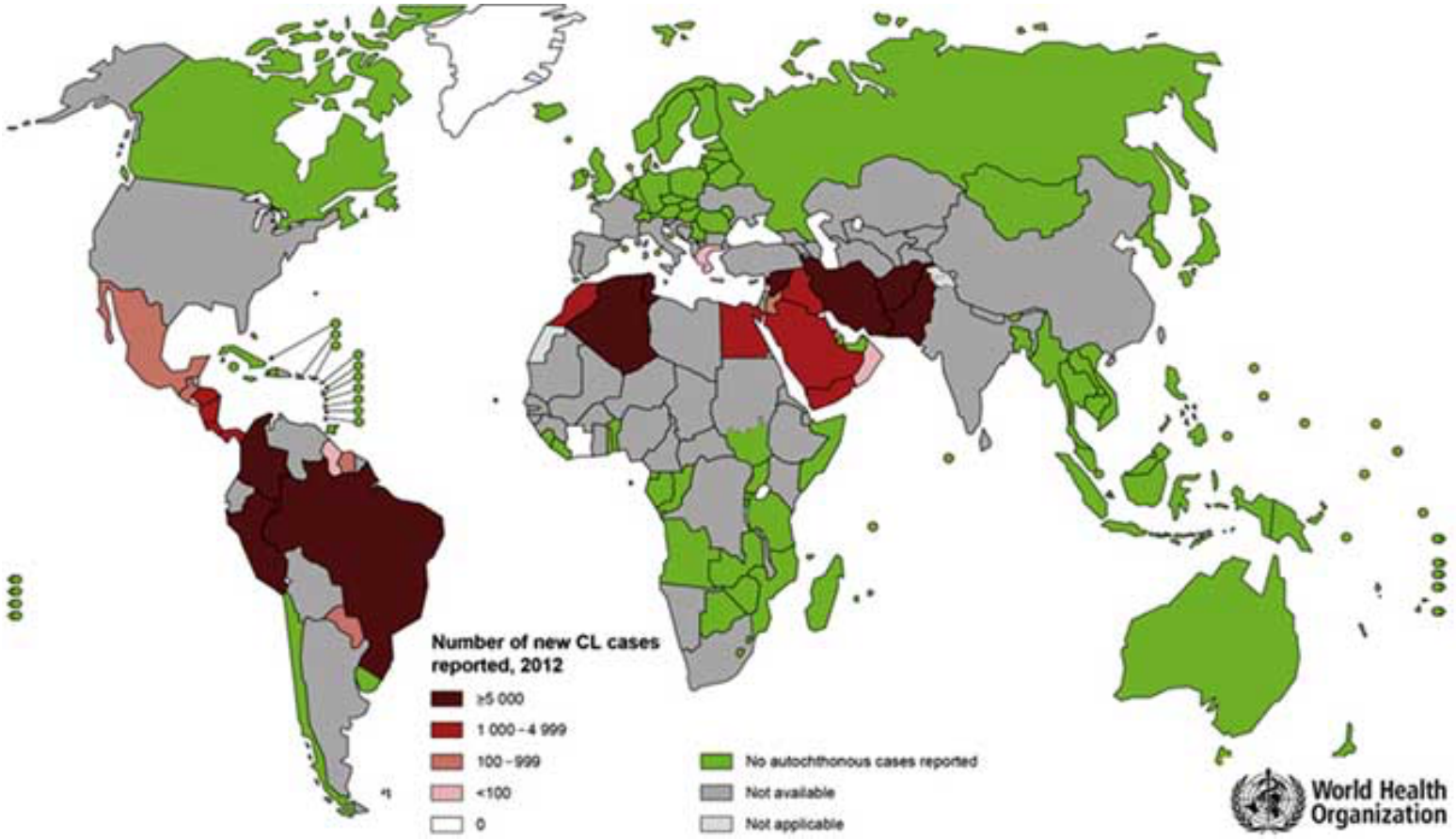
Leishmaniasis : a parasitic disease as an example

- Status of endemicity of **Visceral Leishmaniasis**, worldwide, 2012



Leishmaniasis : a parasitic disease as an example

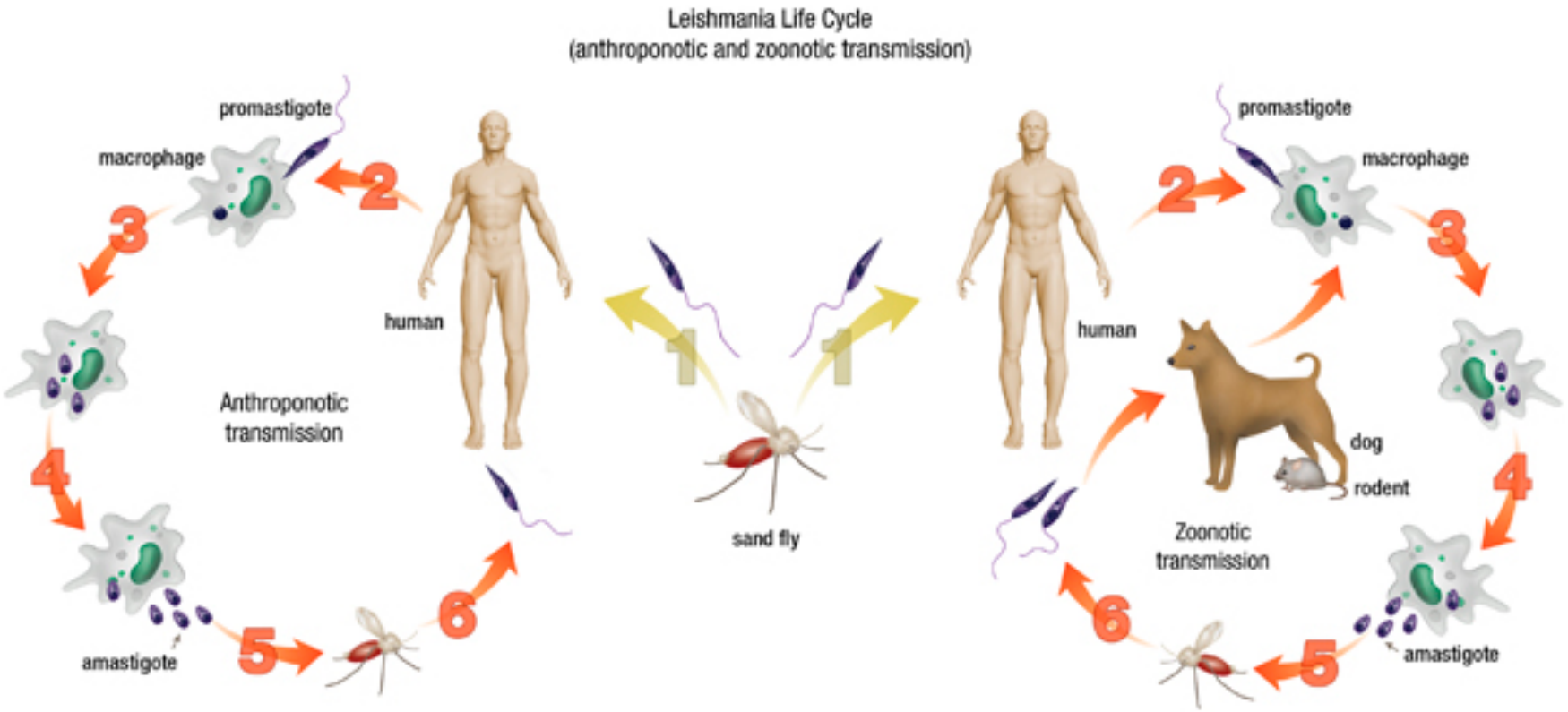
- Status of endemicity of **Cutaneous Leishmaniasis**, worldwide, 2012



About *Leishmania* genome…

Complex life cycle

- Intra- and extra-cellular forms



(Kumar & Engwerda, 2014)

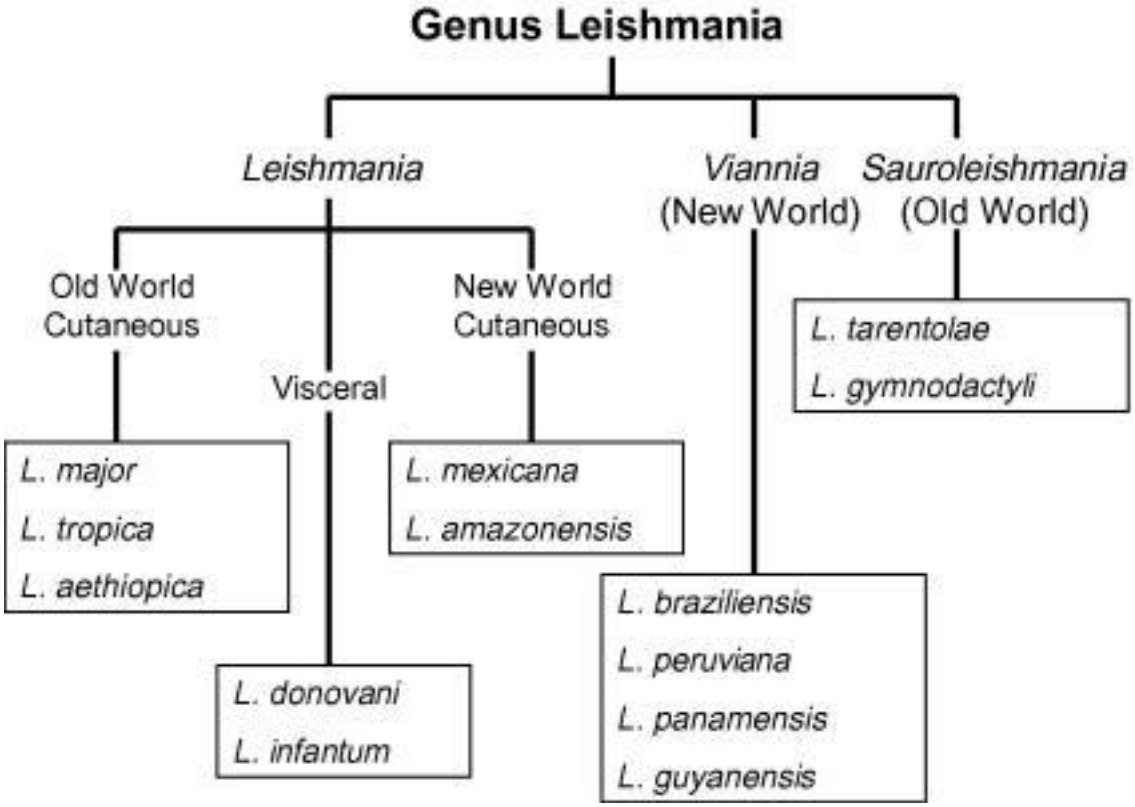
About *Leishmania* genome…

Different species circulate

- New and Old World species

Class Kinetoplastida

Order Trypanosomatida

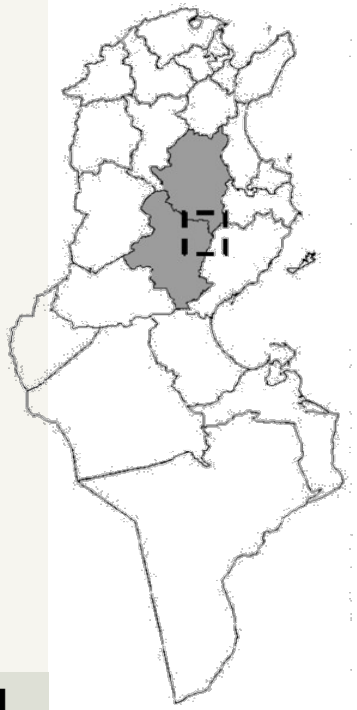


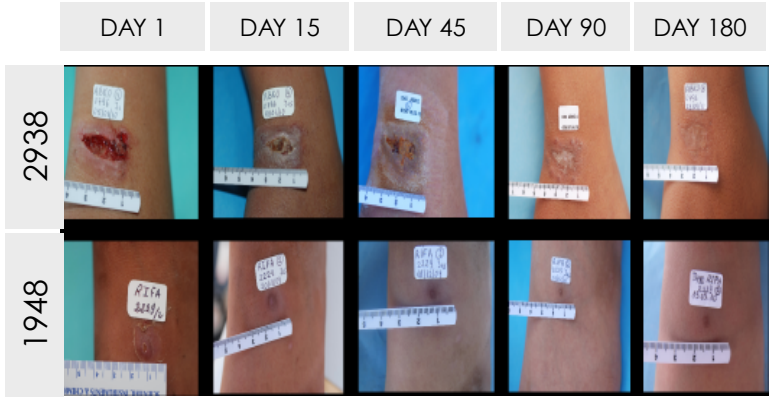
(Bates et al., 2007)
(Hartley et al., 2012)

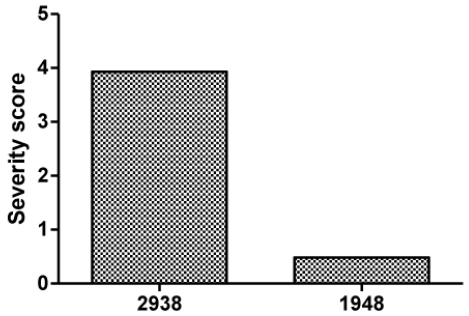
About *Leishmania* genome...

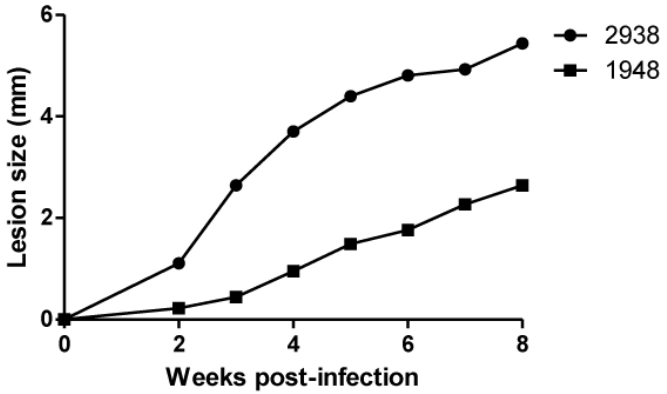
Different species circulate

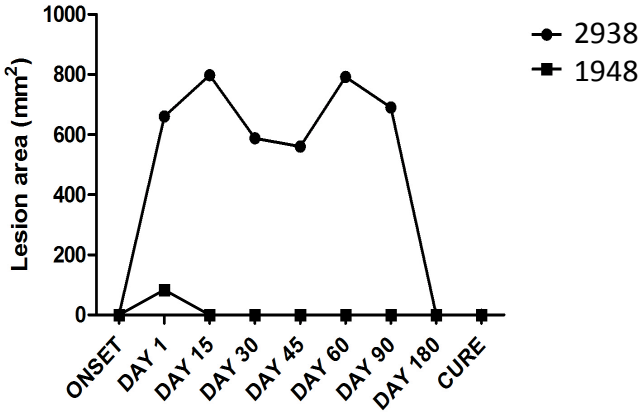
- Intra-species variation is important !!







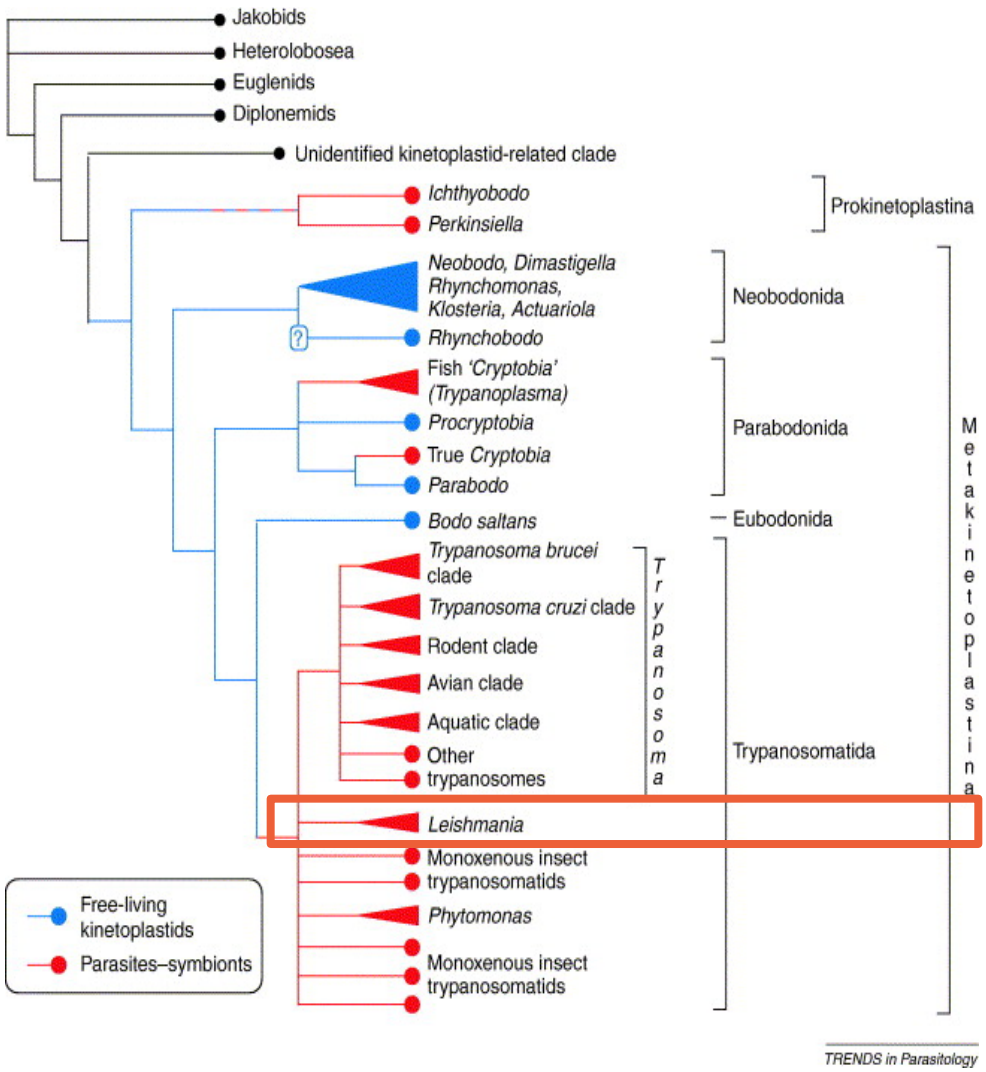




About *Leishmania* genome…

Genome content

- Unicellular eukaryotic organisms
- Mainly known by the disease they cause in humans
- Kinetoplastids
  → flagellated protozoans
- Trypanosomatids
  → various peculiarities

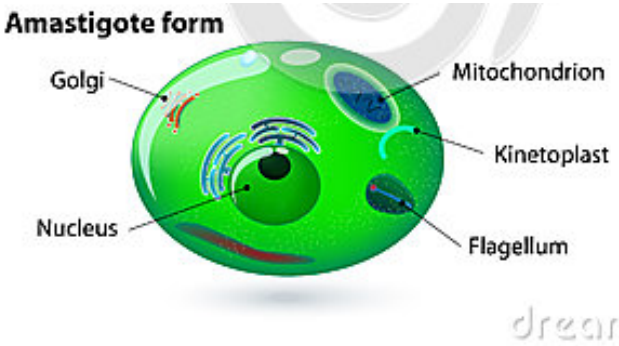


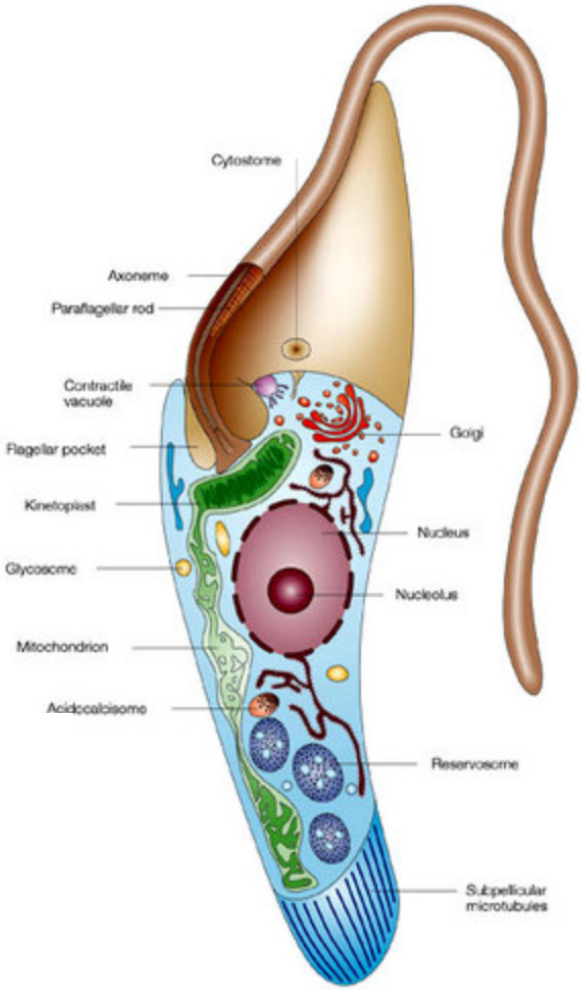
(Simpson et al., 2006)

About *Leishmania* genome…

Genome content

- Nuclear Genome
- Kinetoplastids share a unique mitochondrial DNA structure : kDNA.
- kDNA = network of circular DNA inside a large mitochondrion that contains many copies of the mitochondrial genome.
- kDNA represents ≈ 30% of total DNA



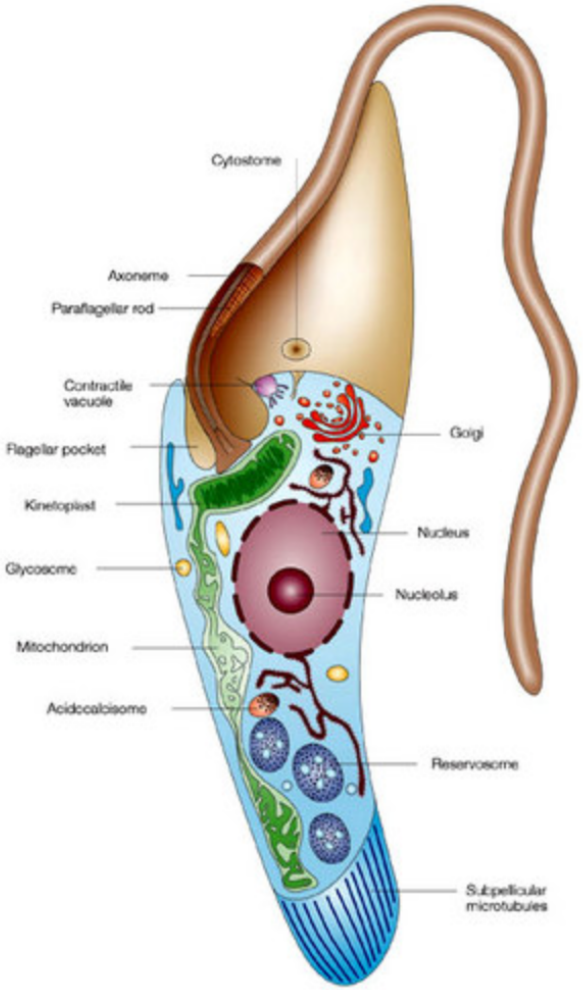


Vargas-Parada, 2010

About *Leishmania* genome…

Genome content

- Unlike other types of DNAs in nature, kDNA is organized in a giant network of interlocked rings : minicircles and maxicircles.
- Maxicircles are between 20 and 40kb in size, up to ≈ 50.
- Minicircles are between 0.5 and 1kb in size, up to ≈ 10,000.
- Maxicircles encode the typical protein products needed for the mitochondria which is encrypted
- The only known function of the minicircles - producing guide RNA (gRNA) = “RNA Editing“.



About *Leishmania* genome…

Genome content

- First *Leishmania* genome chosen to be sequenced as a representative strain: *L. major* MHOM/IL/81/Friedlin.
  (1996-2005, The Leishmania Genome Network used a chromosome-by-chromosome approach shared between SBRI (Seattle, USA), WTSI (UK) and EuLeish consortium).
  (1996-2003 clone-based approach : chr1, 3 & 4 /// 2003-2005 shotgun for chromosomal DNA isolated by PFGE)

- Tryps have relatively small genomes: ≈ 35Mb for *Leishmania*

- The genome consists generally of 34 to 36 chromosome pairs (New/Old World), because fusion of chromosomes could occur
  Numbered ≈ according to their size from 1 (≈250kb) to 36 (≈4Mb).

(Leishmania: After the Genome, P. Myler, chapII)

About *Leishmania* genome…

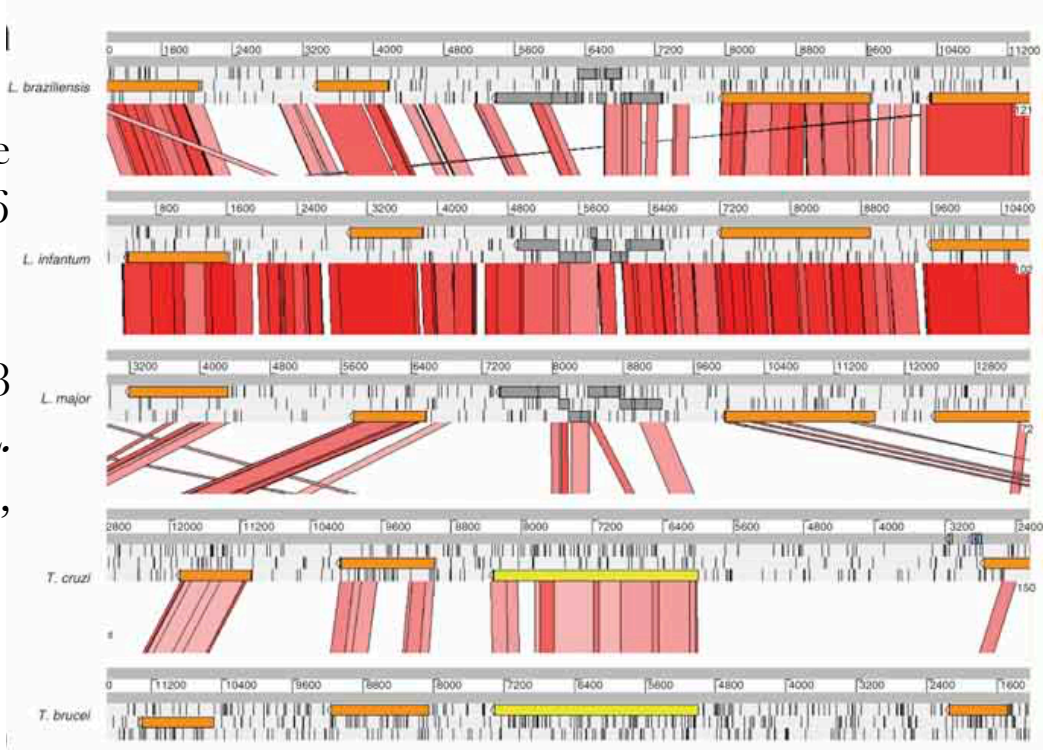
Genome content

- The various *Leishmania* species have very similar genomic arrangements.

→ High degree of Synteny, despite an estimated divergence of 46 million years

Ex:

Comparative genomic analysis of 3 characterized *Leishmania* species: *L. major* Friedlin, *L. infantum* JPCM5, and *L. braziliensis* M2904



- The Genome is GC-rich

The G+C contents of coding (62,5%) and non-coding (57,3%) regions are ≈ equivalent.

(Peacock et al., 2007)
(Lukes et al., 2007)
(P. Myler)

About *Leishmania* genome…

Gene content

- Number of protein-coding genes (Gene prediction + manual curation +…):
  2005 : 8,272
  2007 : 8,378 (include 74 pseudogenes)
  2014 : 8,400 (include 93 pseudogenes) (+ RNA genes = 63 rRNA, 830 small ncRNAs)
  …

- Generally no introns in CDS (rare exceptions)

About *Leishmania* genome…

Gene organization

- One of the most remarkable finding of the Trypanosomatids genome projects : protein-coding genes are arranged into long polycistronic gene clusters / transcription units (PGC/PTU)

- PGC can be organized in a peculiar manner:
  - **head-to-head** (at convergent SSR) 1
  - **tail-to-tail** (at divergent SSR) 2
  - **tail-to-head** 3
  - long clusters of genes can be separated by RNA genes

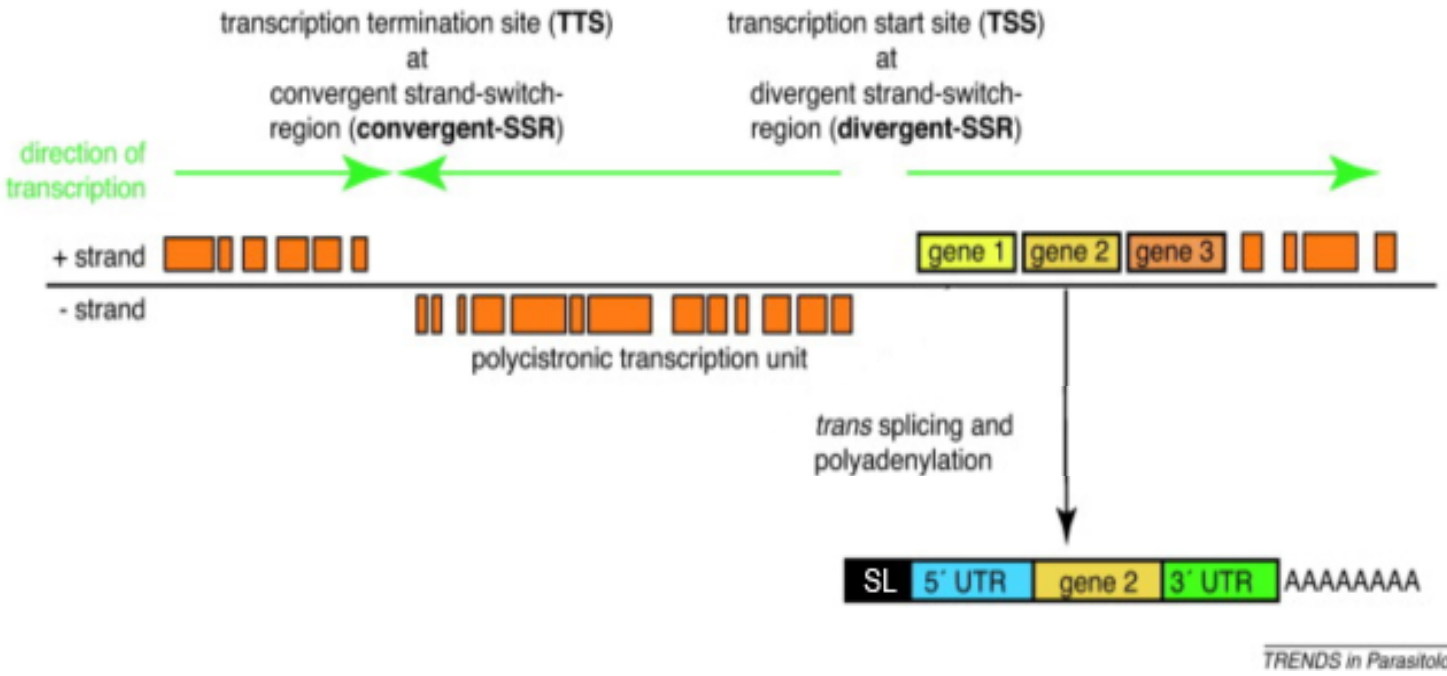
(Tiengwe et al., 2014)

About *Leishmania* genome…

Gene organization

- PGC → up to 100s of genes on the same DNA strand.

100s of PGC, spanning up to several 100s kb.
individual chromosomes may contain several PGC.



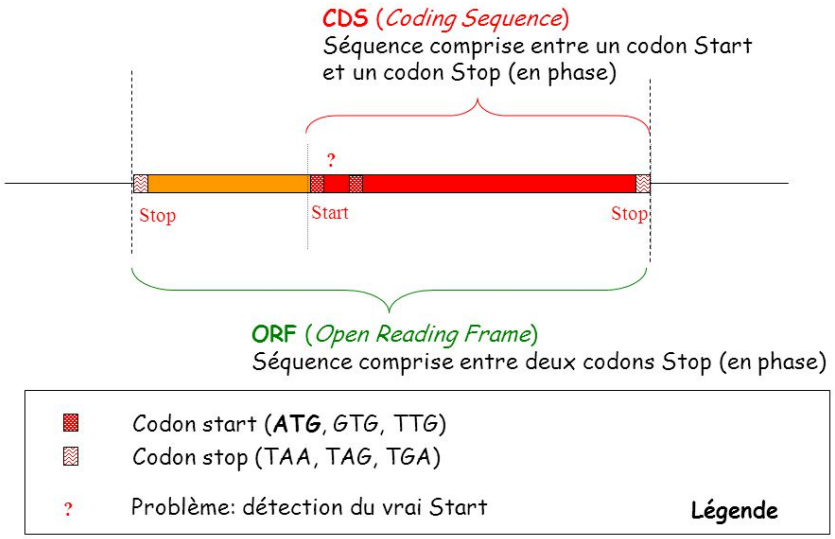
(modified from Siegel et al., 2011)

About *Leishmania* genome…

Gene organization

CDS ≠ ORF

- ORF = Open Reading Frame
  CDS = Coding Sequence



- 5'UTRs and 3'UTRs not clearly defined

- Role of DNA secondary structure signals in gene expression = high DNA curvature was found to be associated with regions implicated in transcription initiation

(Viari et al)
(Smircich et al., 2013)

Complexities of the *Leishmania* genome

Different kind of problems relevant for DNA-Seq analysis

- Lack of a reference genome (ex: isolates)
- Lack of a complete list of functionally annotated genes
- Limited number of species-specific (or isolates-specific) genes
- Multicopy Genes
- Genomic non-Structural Variations (non-SV)
- Genomic Structural Variations (SV)

Complexities of the *Leishmania* genome

Lack of a reference genome

- The completion of the first *Leishmania* reference genome was an important achievement in the quest for genomic information...

- However, it became rapidly clear that strains / isolates from one same species can show geographically-related peculiarities !
→ A reference genome for each ?

phylonetworks.blogspot.com

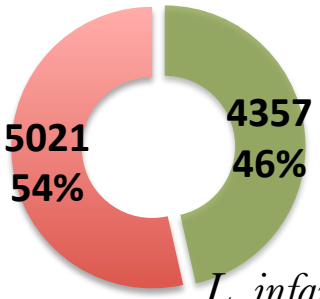
Complexities of the *Leishmania* genome

Lack of a complete list of functionally annotated genes

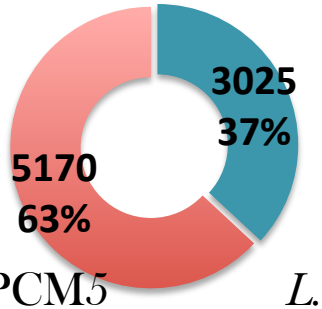
- Gene content: problem of hypothetical genes

Only ≈35-50% genes could be assigned a putative biological function (sequence similarity to proteins in other organisms or experimental characterization...)

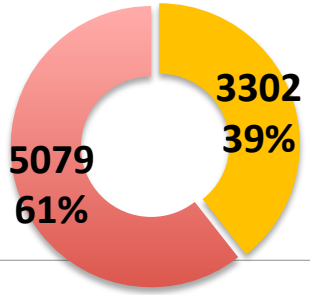
- *L. major* Friedlin



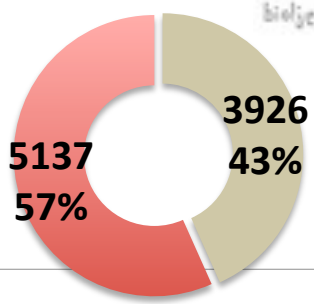
- *L. donovani* BPK282A1



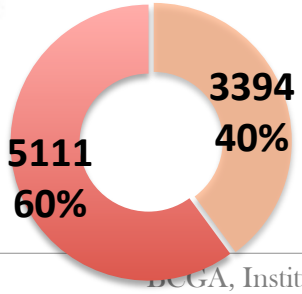
- *L. infantumi* JPCM5



- *L. mexicana*



- *L. Braziliensis M2904*

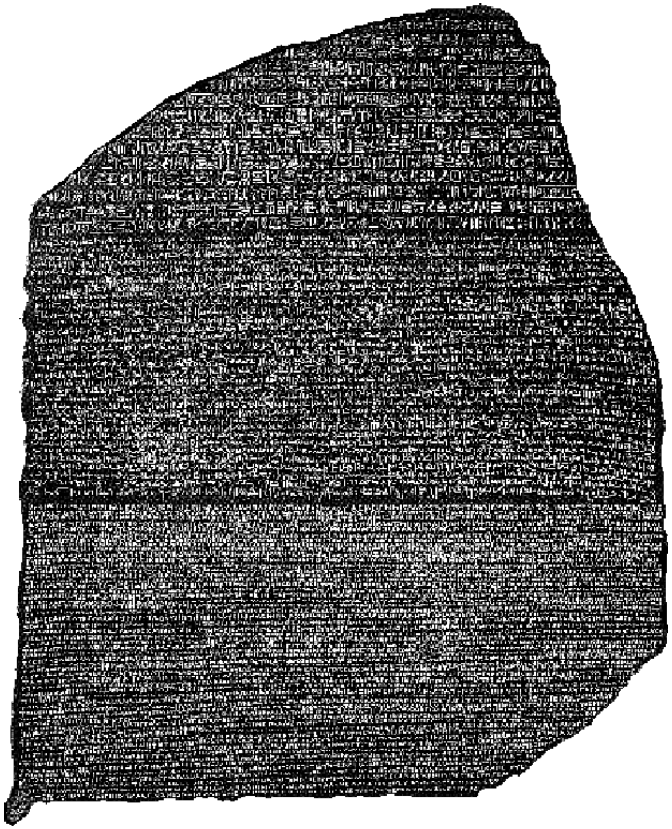


Complexities of the *Leishmania* genome

Lack of a complete list of functionally annotated genes

- Comparative genomics to help

Comparative genomics is based on homology and evolutionary dynamics between organisms.



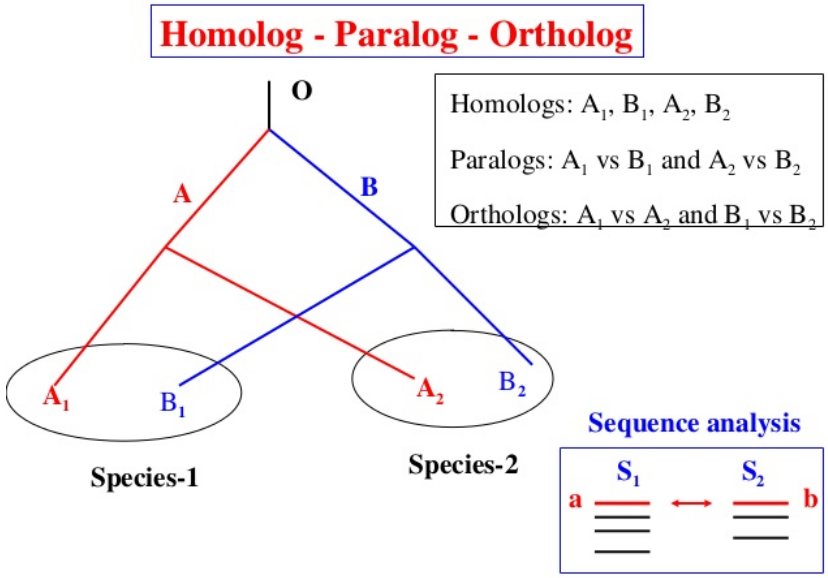
- Gene content: problem of orphan genes

= Genes without detectable homologues in other lineages → are a subset of taxonomically-restricted genes (TRGs), which are unique to a specific taxonomic level (e.g. plant-specific) → usually considered unique to a very narrow taxon, generally a species.

Complexities of the *Leishmania* genome

Lack of a complete list of functionally annotated genes

- Comparative genomics to help
- Data mining using the available related genomes (*Leishmania(s)*, *T. brucei*, *T. cruzi*) provide additional tools for the identification, comparative analysis and functional definition of genes.
- Homologous genes share a common ancestor (intra or inter-species).
- Several scenarios to deal with: orthology, paralogy, horizontal gene transfer, gene loss, orphan genes...



Tekaia, 2014

Complexities of the *Leishmania* genome

Limited number of species-specific (or isolates-specific) genes

- Comparative genomic analysis of 3 characterized *Leishmania* species: *L. major* Friedlin, *L. infantum* JPCM5, and *L. braziliensis* M2904 revealed:

  - unexpectedly small number of species-specific genes. Most of these encode predicted proteins (most of no known function) :

  → proposed to contribute to the parasite tropism and pathology associated with the different forms of leishmaniasis.

- Proof: *L. donovani*–specific genes expressed in *L. major* showed a significant increase in parasite survival in visceral organs in mice → individual genes can contribute to parasite tropism in the host.

(Peacock et al. 2007)
(Smith et al., 2007)
(Zhang et al. 2008)
(Zhang and Matlashewski 2010)

Complexities of the *Leishmania* genome

Multicopy genes

- **Multicopy genes (tandem arrays)** can generally be defined as genes of more than one copy that belong to the same orthologous group and are encoded on the same chromosome.

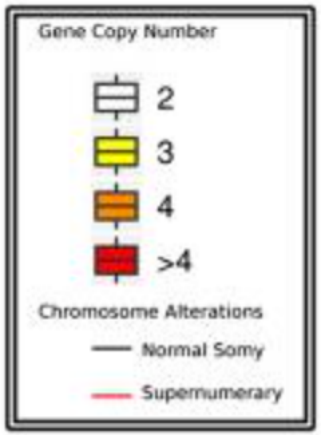
- The **highly repetitive nature** of **tandem arrays** is problematic for *de novo* genome assembly, leading to "collapsed" arrays of unknown length.

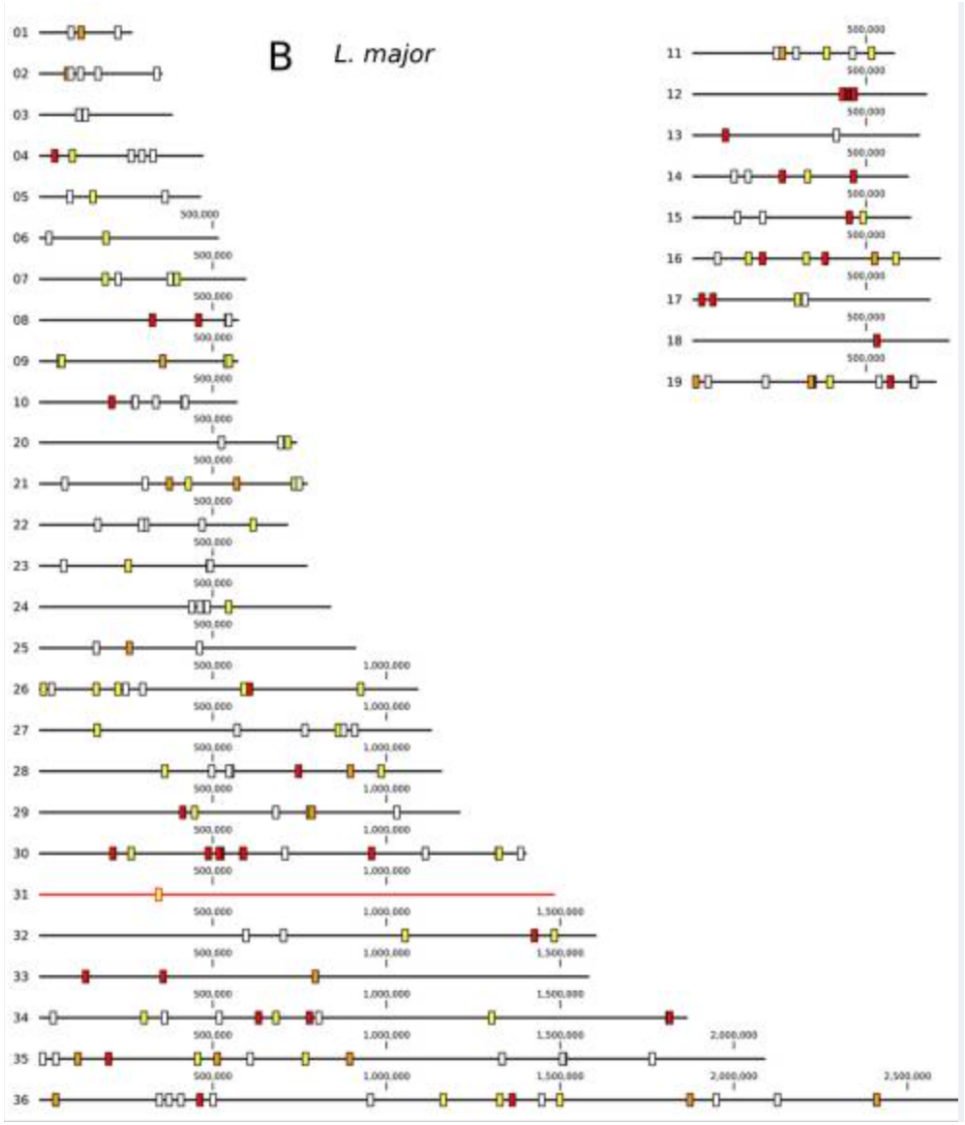
(Rogers et al., 2011)
(Ivens et al. 2005)

Complexities of the *Leishmania* genome

Multicopy genes

Distributions of multicopy genes on each chromosome in *L. major* Friedlin.





(Rogers et al., 2011)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

- Structural variation is the variation in structure of an organism's chromosome. It involves rearrangements in the order of sections of DNA.

- It consists of many kinds of variations:
  - deletions
  - duplications
  - copy-number variants
  - insertions
  - inversions
  - translocations (cryptic or not), inter or intra-chromosomal

Can also be complex, involving multiple events at the same location.

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

- Challenges in defining SVs :

-> Terminology and classification !!!

There is no standard nomenclature for structural variants !

Can be classified according to the cytogenetic or mutation literature (for example, indels).

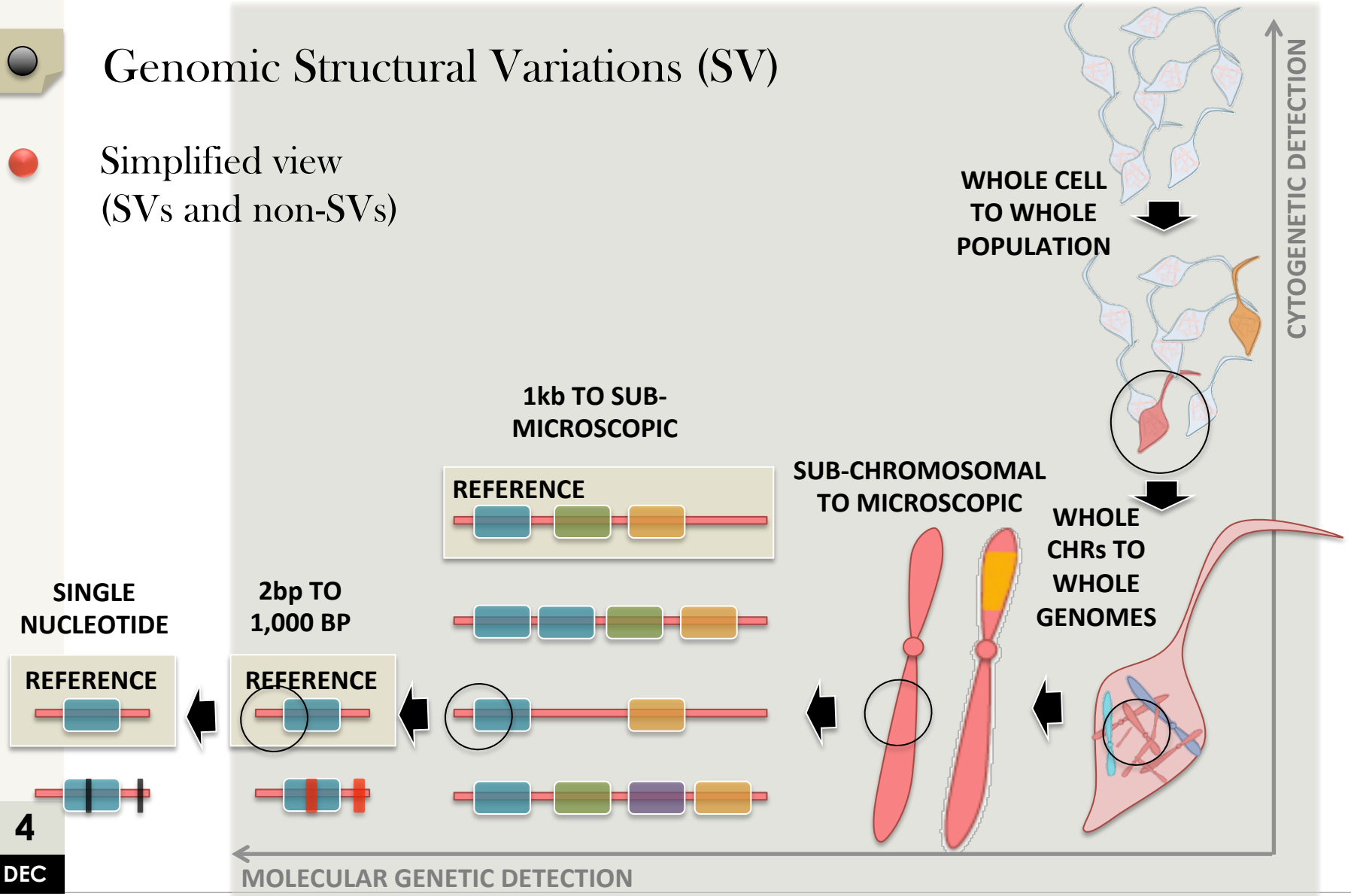
For some terms, such as CNV, there is added complication because different classes exist.

(Scherer et al., 2007)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

Simplified view
(SVs and non-SVs)



Fatma Guerfali, PhD
(inspired from Scherer et al., 2007)
BCGA, Institut Pasteur de Tunis

Complexities of the *Leishmania* genome

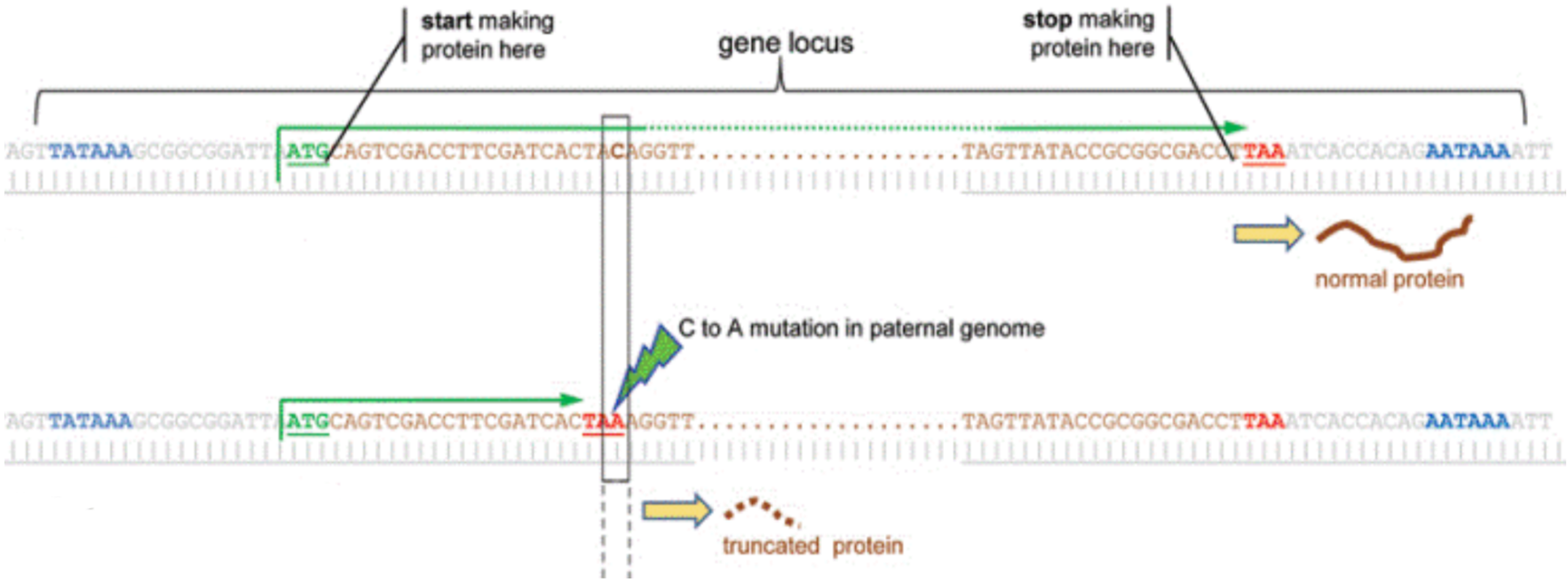
Genomic non-Structural Variations (SV)
Single Nucleotide Polymorphism (SNPs)

- When a change in the DNA is created it is called a mutation.
- Generally, if it is passed on to future generations and becomes established in part of the population we refer to it as a **polymorphism**
  = a difference in the genetic code present in some people at a specific position in their genomes.
- Small pieces of DNA may be lost or inserted (**InDels**), but the most common form of polymorphism is the alteration of a single base = **Single Nucleotide** Polymorphisms, or SNPs.

Complexities of the *Leishmania* genome

Genomic non-Structural Variations (SV)

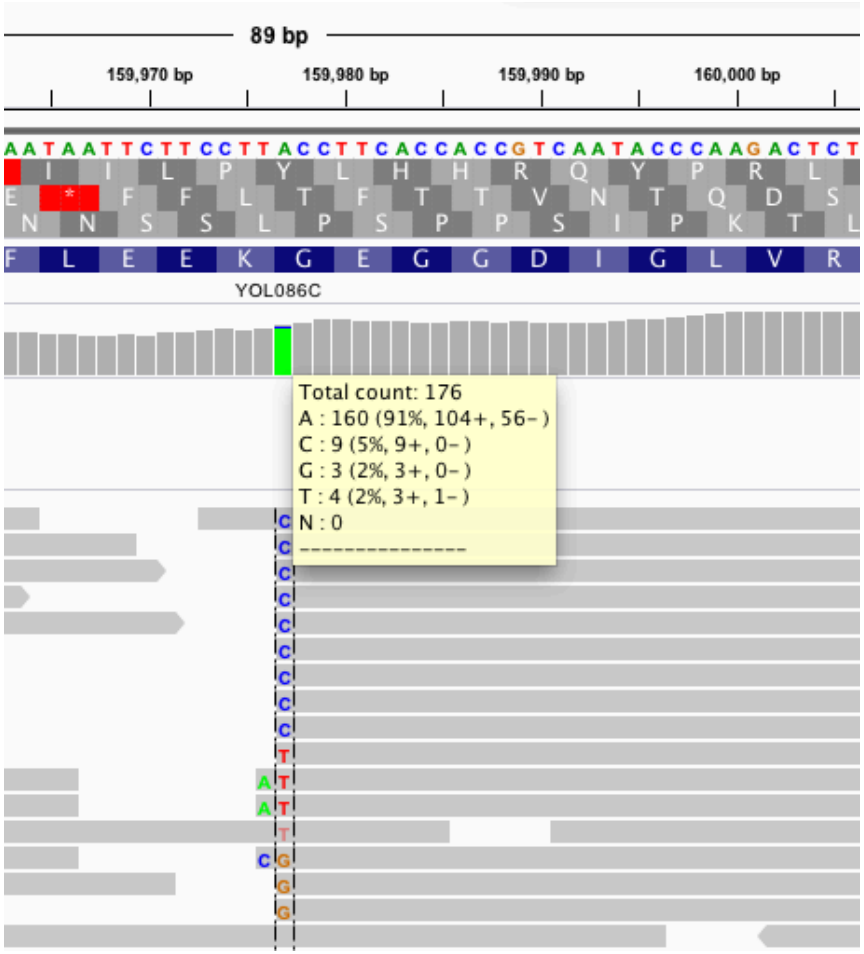
Single Nucleotide Polymorphism (SNPs)



Complexities of the *Leishmania* genome

Genomic non-Structural Variations (SV)

Single Nucleotide Polymorphism (SNPs)



(www.broadinstitute.org)

Complexities of the *Leishmania* genome

Genomic non-Structural Variations (SV)

Single Nucleotide Polymorphism (SNPs)

- The use of NGS for drafting many genomes can now quickly uncover **SNPs, insertions and deletions** by mapping reads against a well-annotated **reference genome**.
→ provide a list of possible differences that may be the basis for any functional variation among strains.

- The importance of SNPs is particularly highlighted by the fact that many pathogens are closely related, so SNPs in conserved core genes have been used to discriminate and infer phylogenetic relationships between closely related pathogenic strains.

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

Size 2bp – 1,000bp

- The genome of *Leishmania* was found to be relatively rich in microsatellites.
  ex: about 600 (CA)n loci per haploid genome.
- Microsatellite = simple sequence repeats (SSRs) or short tandem repeats (STRs)
  → repeated motifs of 1–6 nucleotides found in all euk and prok genomes.
- Microsatellites mutate at orders of magnitude higher than the bulk of DNA
  → particularly useful for studying variation between closely related organisms.
- Microsatellite sequence variation results from the gain and loss of repeat units.
- The evolutionary history of a particular repeat sequence may be uncertain
  → panel of 10–20 unlinked microsatellite markers.

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

Size 2bp – 1,000bp

- Variations in the copy number of DNA segments account for a substantial amount of genome diversity of most organisms.

- DNA amplification in the parasite *Leishmania* :

  - is a contributor to copy number variation.

  - can occur in response to various stresses or after altered growth conditions.

  - arises by DNA rearrangements involving homologous repeated sequences = **homologous recombination (HR) between direct repeated sequences (DRs).**

  - **large inverted duplications are generated by the annealing of IRs followed by duplications.**

(Ouellette et al., 2014)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

Size 2bp – 1,000bp

IRs :

- **Are** widespread in the *Leishmania* genome
- are known to increase chromosome instability during replication, representing a substantial **source of DNA breakage and rearrangement.**

→ most of the ***Leishmania*** **genome is subjected to stochastic gene rearrangements mediated by these low-copy repeat sequences.**

(Ouellette et al., 2014)

Complexities of the *Leishmania* genome

- Genomic Structural Variations (SV)
  Size 2bp – 1,000bp
- Homologous repeated sequences are widespread in the *Leishmania* genome.
- Repeated sequences used for DNA amplification are generally noncoding and are interestingly highly conserved between different *Leishmania* species.
- These intergenic sequences may have been maintained to facilitate the amplification of key genomic loci essential to respond to changing growth conditions

The distances between DRs or IRs are on average between 1 and 100 kb.
IRs are found in general closer to telomeres, whereas DRs appear more evenly distributed along the chromosomes.

(Ouellette et al., 2014)

Complexities of the *Leishmania* genome

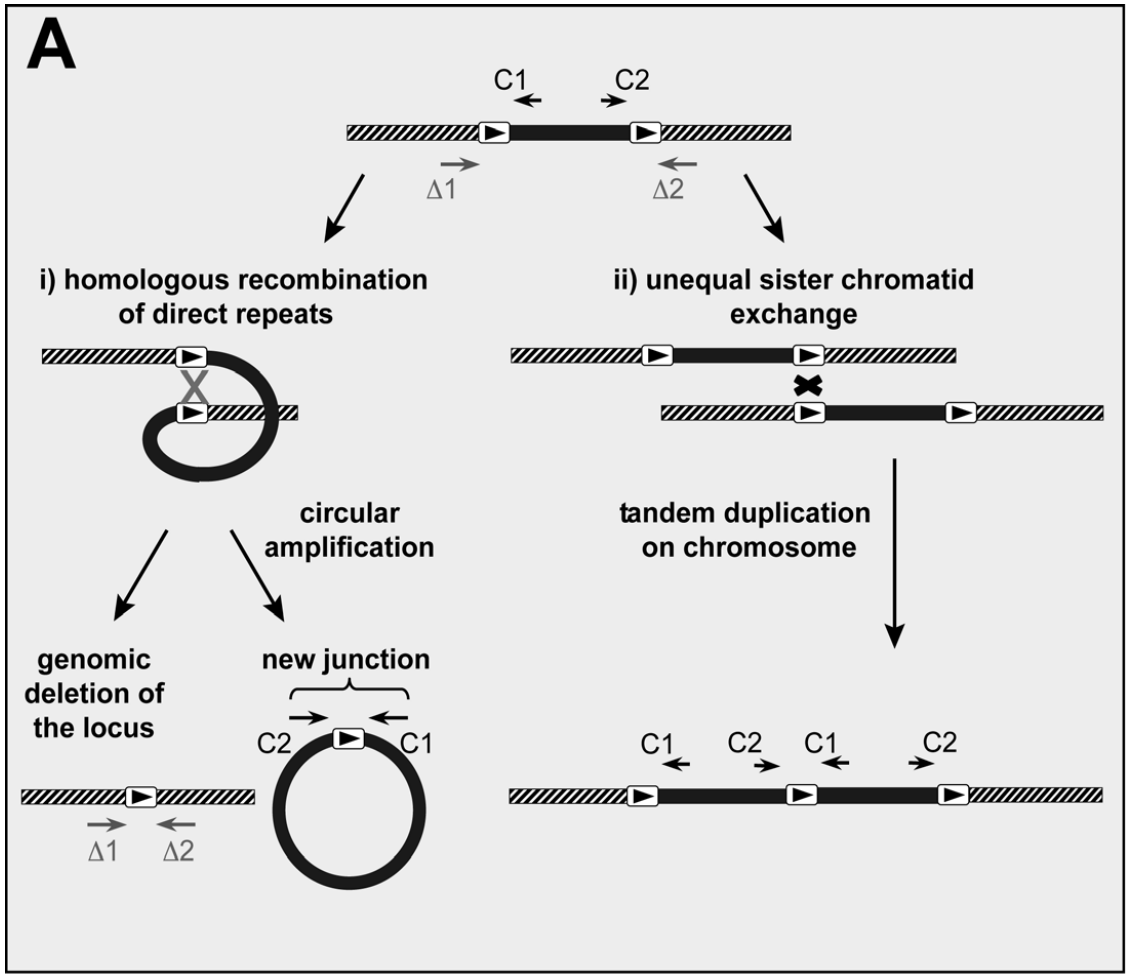
- Genomic Structural Variations (SV)
  Size 2bp – 1,000bp

- **Example of Gene amplification in *Leishmania* : DRs**

HR between DRs can lead to:

(i) circular amplification (conservative & non-conservative)

(ii) tandem duplication of the locus by nonequal crossing at sister chromatids.



(Ouellette et al., 2014)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

Size 2bp – 1,000bp

- Repeated sequences clustered into RAGs (Repeat Alignment Groups) ≈ 500.
(1 RAG = all the members of a same repeat family)

- Within RAGs, detection of SIDERs (sequences Short Interspersed DEgenerate Retroposons).

SIDER1 or SIDER2 subfamilies : functional + structural roles

- regulation of gene expression (post-transcriptional (SIDER2) or translational (SIDER1) levels)
- participation to recombinational events leading to genetic amplification.

- TATEs: Telomere-Associated-Mobile-Element.
Unlike SIDERs, are putatively active mobile elements. Located first in telomeres, but found also in internal positions in chromosomes.

Complexities of the *Leishmania* genome

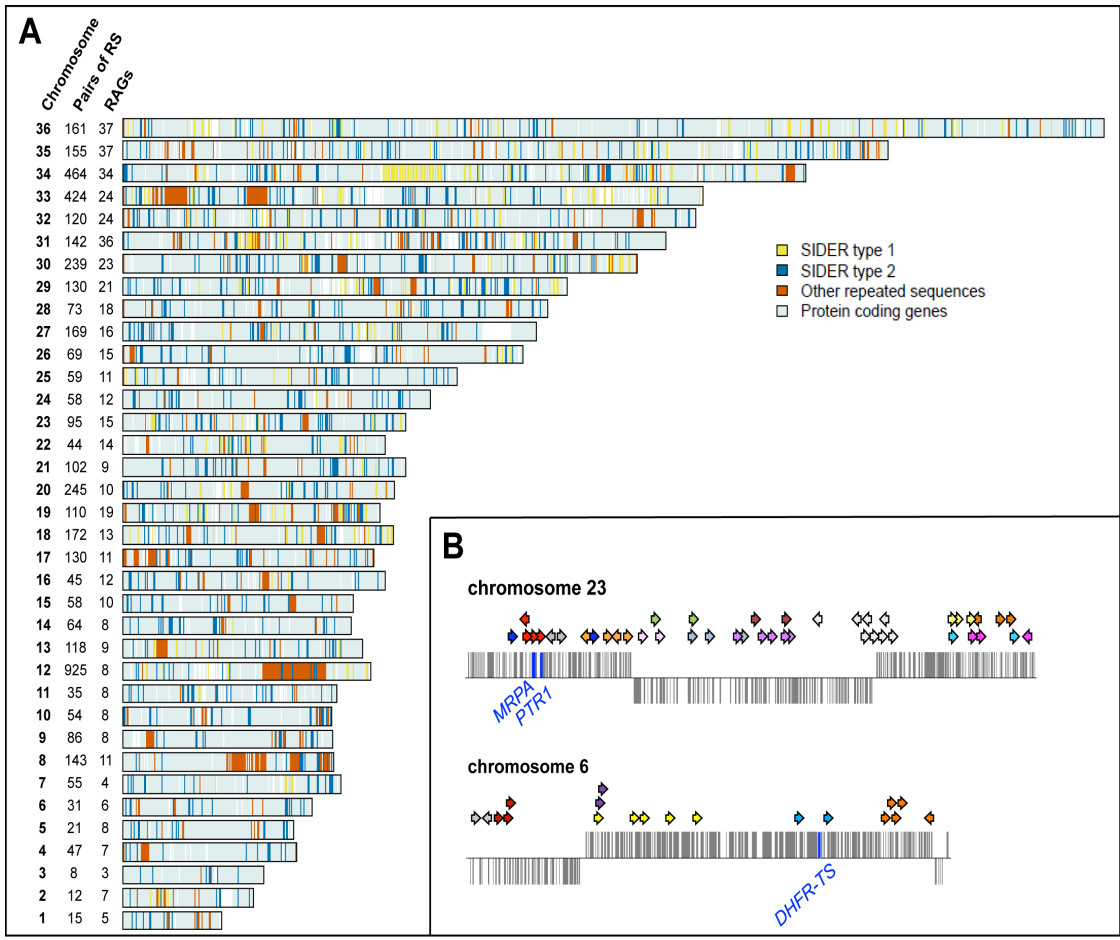
Genomic Structural Variations (SV)

Size 2bp – 1,000bp

- The repeats are part of distinct RAGs.

(A) Repeats belonging to the SIDER family are in yellow (SIDER1) or blue bars (SIDER2).
Other repeated sequences are represented by orange bars.

(B) Repeats on chromosomes 6 and 23 of *L. major*. Repeats of the same RAG = same color.



(Ouellette et al., 2014)

Complexities of the *Leishmania* genome

- Genomic Structural Variations (SV)
  1kb to submicroscopic (and more) = CNVs

- Copy number variation (CNV) is defined as the gain or loss of genomic material, **usually more than 1kb.**

- May have a phenotypic impact by altering the fitness of an organism.

- CNV creates paralogous genes that may evolve differently than the progenitor gene or that may alter the expression level of a gene or genomic region.

(Reis-Cunha et al., 2015)

Complexities of the *Leishmania* genome

- Genomic Structural Variations (SV)
  1kb to submicroscopic (and more)

- Copy-number variation (CNV) is a large category of structural variation, which "could" include insertions, deletions and duplications, although they are generally separated.

- A CNV is having more or less than the expected 2 copies of a region of DNA.

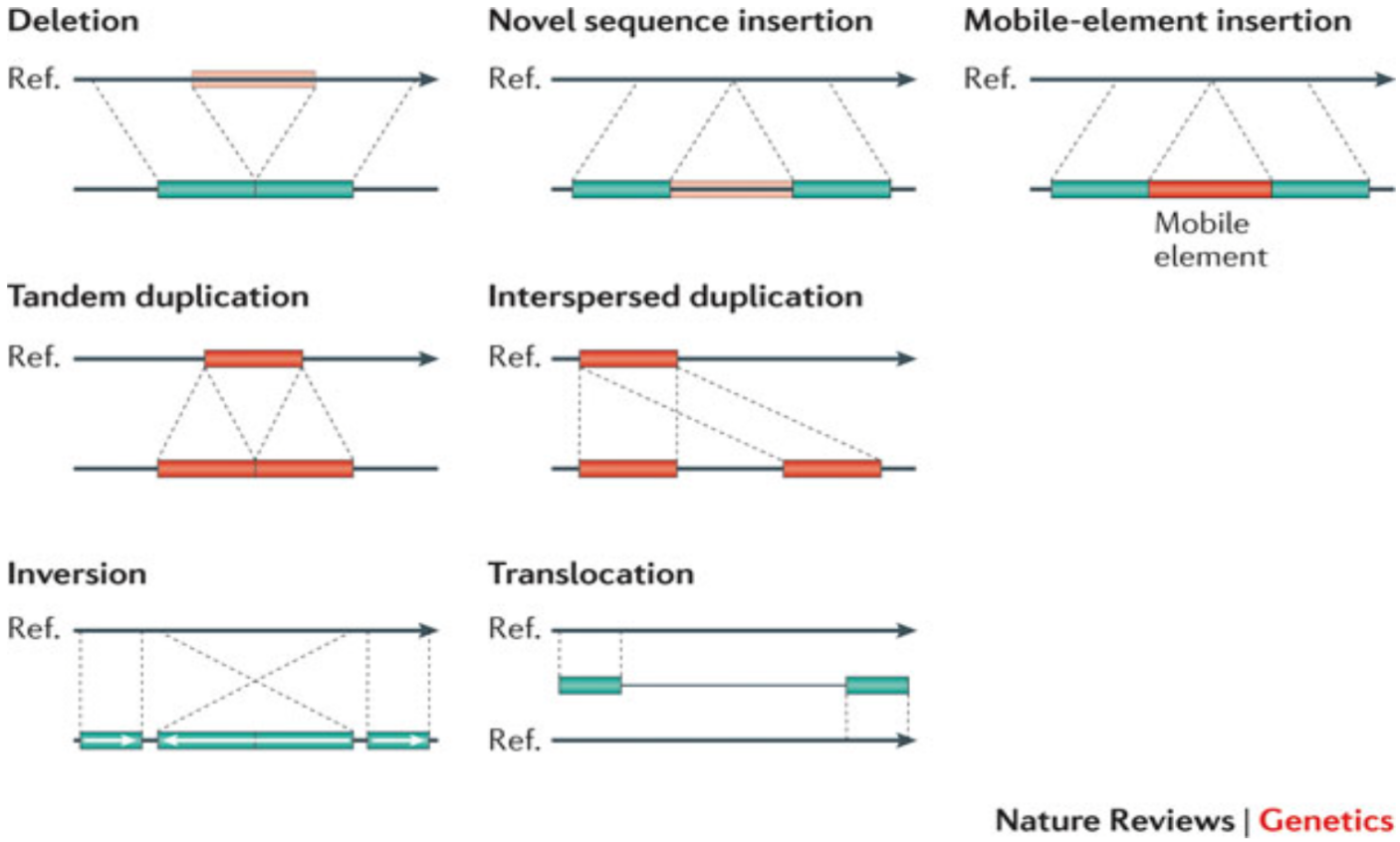
- CNVs in human genome affect more nucleotides than Single Nucleotide Polymorphism (SNP)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

1kb to submicroscopic (and more) = CNVs

Described as structural changes in regards to the reference genome



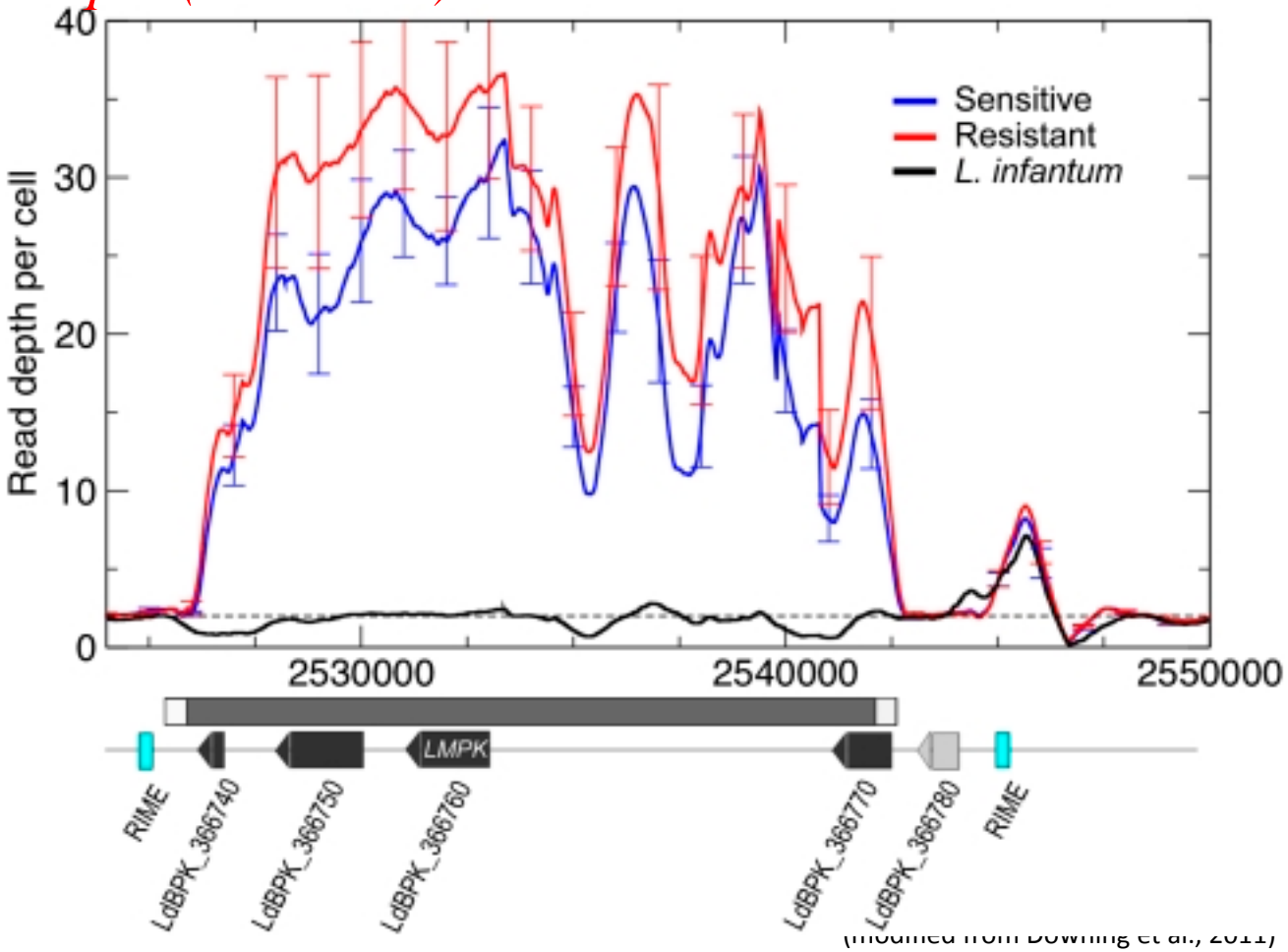
Nature Reviews | Genetics

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

1kb to submicroscopic (and more) = CNVs

Example of CNV



(modified from Downing et al., 2011)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

1kb to submicroscopic (and more) = CNVs

How to define copy number: 2 different metrics:

- **Haploid number** = number of genes on an individual chromosome.
- **Gene dose** = total copy number of a particular gene in the genome of a cell, accounting for copy number of the chromosome.

Ex: a single-copy gene in a diploid organism has :
Haploid number = 1 ; Gene dose = 2.

Gene ID	Calculated Copy Number	Estimated Haploid Number	Chr Ploidy	Estimated Gene Dose
LmjF.12.0730	25.88	26	2	52

(Sterkers et al., 2014)
(TriTrypDB.org)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)
Chromosome/Cell/Population

- *Leishmania* genomes are **generally diploïd.**

- There were remarkably very few (≈0.1%) sequence differences between homologous chromosomes in the reference Friedlin genome.

- However:
  - varying degrees of aneuploidy can occur
    - Whole chromosome aneuploidy
    - Partial aneuploidy
    - Mosaïc aneuploidy
  - DNA amplifications are frequent under nutritional stress or drug selection

(Leishmania: After the Genome, P. Myler, chapII)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)
Chromosome/Cell/Population

- ***Aneuploïdy :***

"describes a chromosome number which is not an exact multiple of the haploid number. "

x = number of unique chromosomes
Multiples of x = euploïds
(2x = diploïds, 3x= triploïds, ..., polyploïds)
Any modification from the number x = aneuploïds

- Aneuploidy is generally considered as '**whole-chromosome**' aneuploidy "as opposed to **copy-number changes** affecting only parts of chromosomes, which are described as 'partial', 'segmental', or 'structural' aneuploidy."

(Sterkers et al., 2014)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)
Chromosome/Cell/Population

Chromosome copy number variation observed in *Leishmania*

Changes in **ploidy** = at the level of whole chromosomes = changes in chromosome copy number

Have been reported in various *Leishmania* species :

- Due to manipulation of essential genes
- during in vitro growth + variation in conditions
- after genetic exchange
- following drug selection *in vitro*.
- Methods of analysis (individual cell vs. cell populations).

(modified from Downing et al., 2011)
(Cruz et al. 1993; Hassan et al. 2001)
(Martinez-Calvillo et al. 2005)
(Akopyants et al. 2009)
(Leprohon et al. 2009)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)
Chromosome/Cell/Population

- **Chromosome copy number variation observed in *Leishmania***

- *L. major* : **disomy for most chromosomes is common**
but some show trisomy or more, perhaps due to

- Chromosome 31 : the only chromosome reported as being supernumerary in **all *Leishmania* species** and isolates analyzed to date.

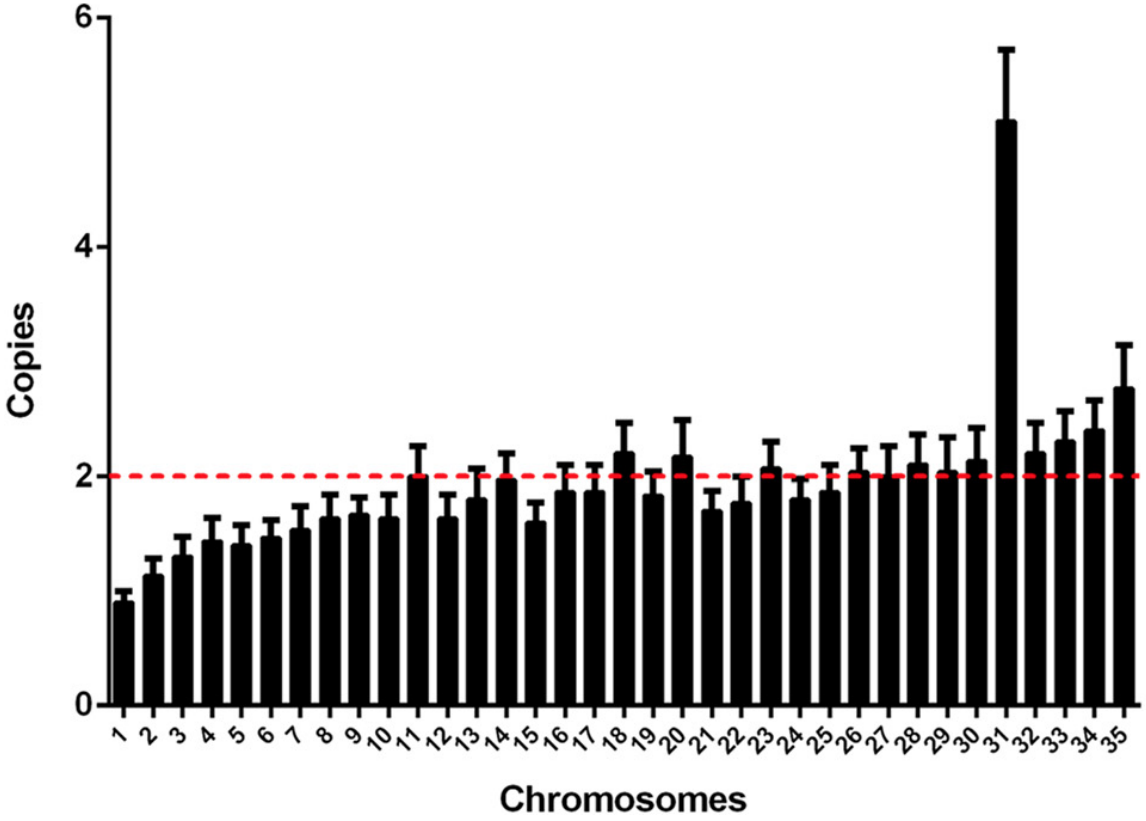
(modified from Downing et al., 2011)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)

Chromosome/Cell/Population

Chromosome copy number variation observed in *Leishmania*



(modified from Downing et al., 2011)

Complexities of the *Leishmania* genome

Genomic Structural Variations (SV)
Chromosome/Cell/Population

- Irregular values for median read depth mainly in smallest chromosomes
= suggestive of chromosomes that are not fully disomic, trisomic...

- One important feature of the massively parallel sequencing achieved with the Illumina analyzer :
= it provides unprecedented read depth coverage across all *Leishmania* chromosomes.

→ Analysis of median read depth for the *Leishmania* genome shows which chromosomes will have an even read depth, indicating that the chromosomes within the population of cells are **disomic or not, fully or not**.

(Rogers et al., 2011)
(Sterkers et al., 2014)

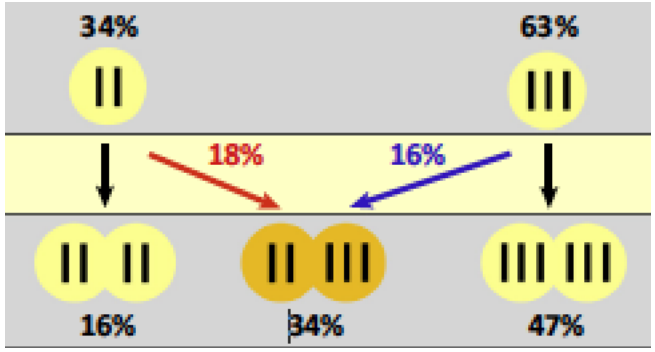
Complexities of the *Leishmania* genome

- Genomic Structural Variations (SV)
  Chromosome/Cell/Population

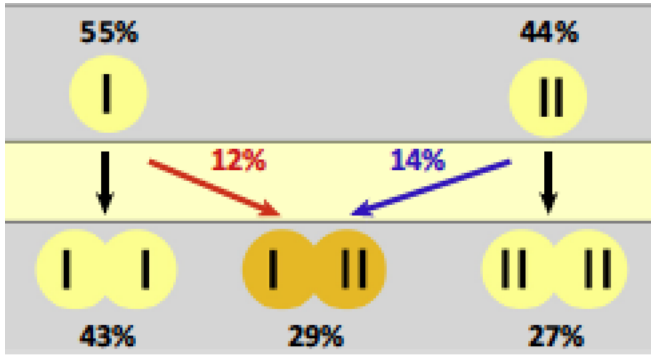
- **Mosaic aneuploidy** = a phenomenon characterized by a variation in the somy of the same chromosomes among cells within the same population / strain.

- FISH (Fluorescence in situ Hybridization) supported by HTS data :
  → highly variable chromosomal contents among cells.

Asymmetric '2+3' divisions generates mosaic aneuploidy of chromosome 5



Asymmetric '1+2' divisions generates mosaic aneuploidy of chromosome 2



(Rogers et al., 2011)
(Sterkers et al., 2014)

Complexities of the *Leishmania* genome

Conclusions

- Most of the *Leishmania* genome is stochastically subjected to **genetic rearrangements** at the level of the low-copy repeated sequences or others (mutations, recombinations, duplications, rearrangements...).

  Modern evolutionary synthesis:
  Evolution is gradual: small genetic changes regulated by natural selection accumulate over long periods.

- **Many structural variants have been associated with genetic diseases, however more are not !!!** Their roles are not completely understood, so a particular care should be given in interpretating their existence or not.

(Ouellette et al., 2014)

Complexities of the *Leishmania* genome

Conclusions

Be carefull in interpreting your results :

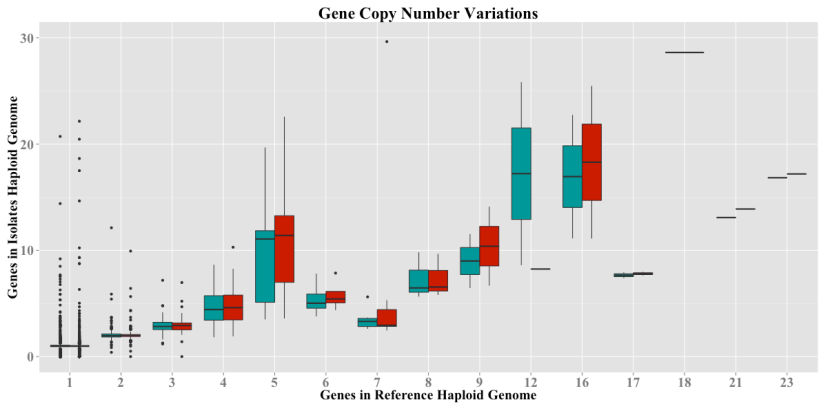
Not all variants are interesting !!!

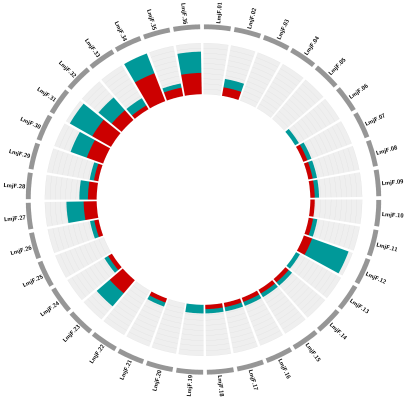
- CNVs are not always related to disease
- Synonymous SNPs have been shown to have effects

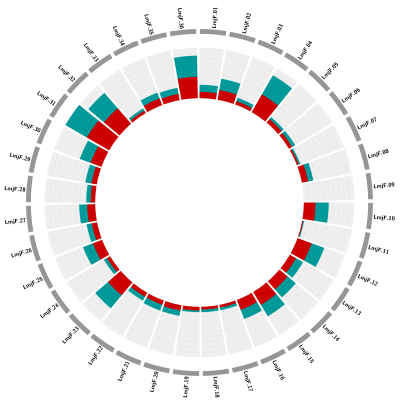
GENE CNVs

HIGH IMPACT SNPs

HIGH IMPACT InDels



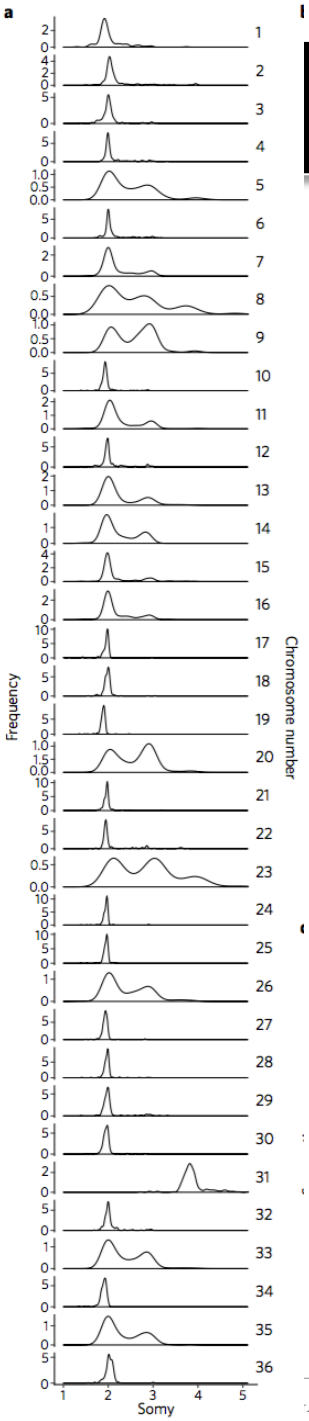




(Ghouila et al., 2016)

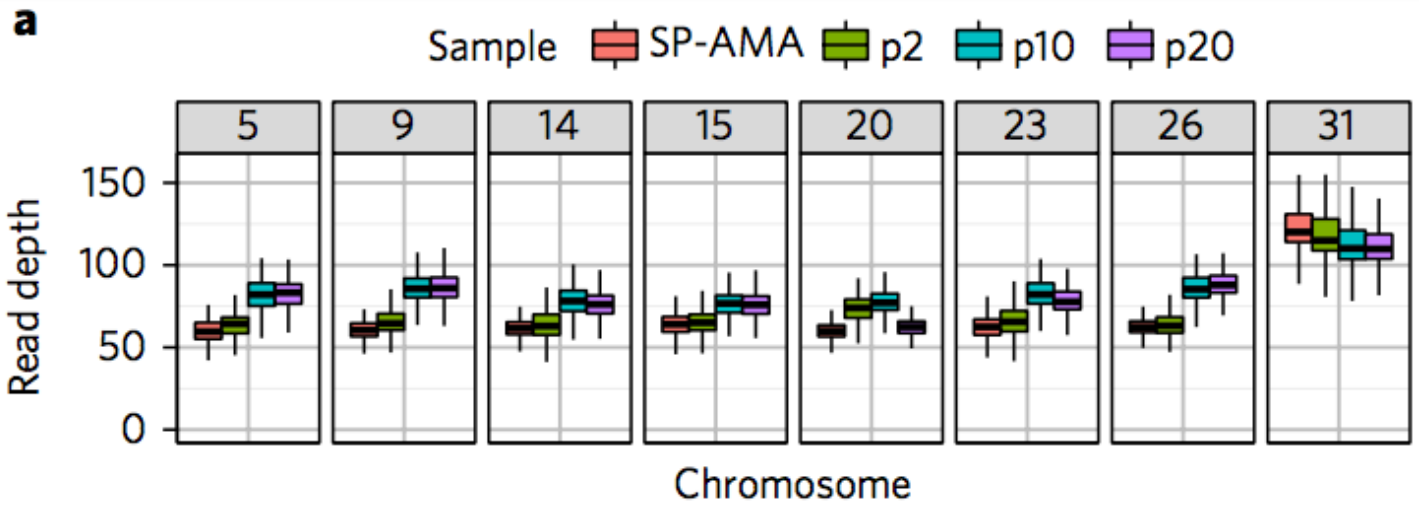
Complexities of the *Leishmania* genome

Polysomy analysis.
Polysomy level was estimated for each chromosome and sample by read-depth analysis. The distribution of chromosome copy number is shown across all field isolates.



(Prieto Barja et al., 2017)

Complexities of the *Leishmania* genome



Follow-up of hamster-derived LD1S amastigotes during adaptation to in vitro culture.

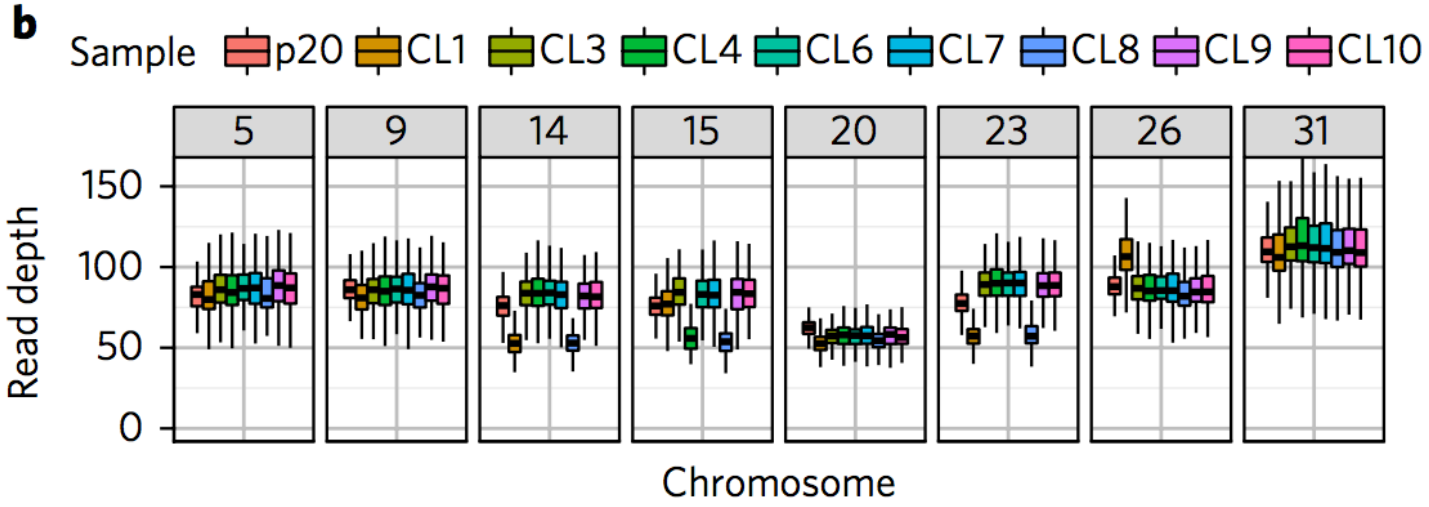
- **parasites rapidly establish stable trisomies** for chromosomes 5, 9, 23 and 26 between in vitro p2 (20 generations) and p10 (60 generations)

→ matches the most common variations observed in the field isolates and they were highly reproducible across two independent experiments.

- **Not all aneuploidies are stable and homogenous, however.** Example:
  - chr20: underwent a transient trisomy between p2 and p20 (190 generations)
  - chr14 and chr15: probable mosaic aneuploidies occurred and were stably maintained as judged by their intermediate read-depth levels at p10 and p20.

(Prieto Barja et al., 2017)

Complexities of the *Leishmania* genome



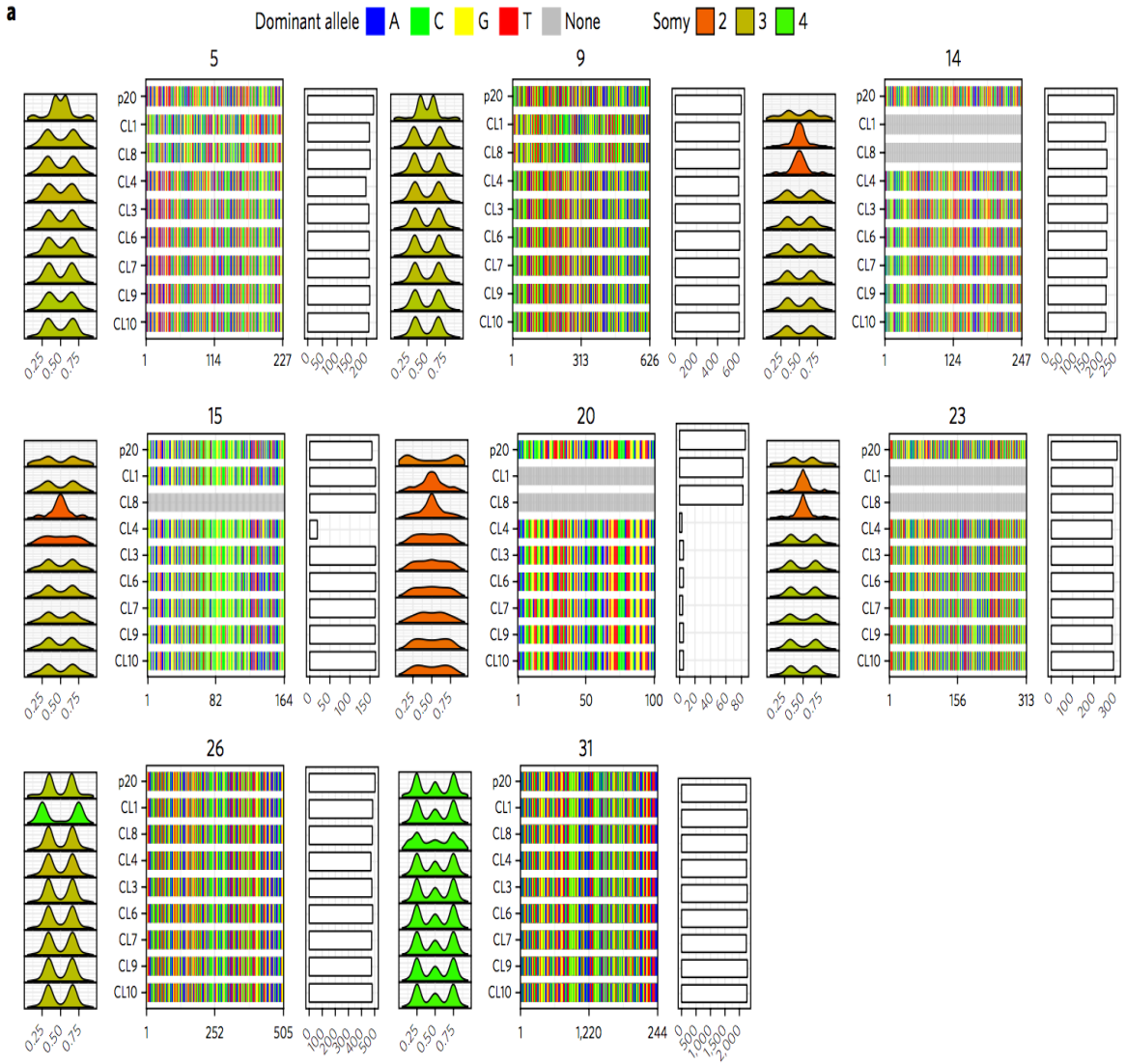
Systematic comparison between p20 and 8 individual subclones to model the original population complexity.

Aneuploidies provide a powerful insight into haplotype phasing, due to systematic frequency shifts associated with duplicated homologue chromosomes.

Additional haplotype and karyotype comparisons suggest that the 8 clones may have arisen from at least two independent founding individuals.

(Prieto Barja et al., 2017)

Complexities of the *Leishmania* genome



Fluctuations of allele frequency during culture adaptation.

Haplotype selection in clones.

- Variable sites in chromosomes undergoing amplification were either coloured according to the major alleles (that is, highest frequency) or left in grey for balanced heterozygote sites.
- The number of heterozygous sites for each chromosome and sample are shown by the histograms on the right. Samples with low counts as observed for chromosomes 15 and 20 represent loss of heterozygosity occurring during culture adaptation.

(Prieto Barja et al., 2017)

Complexities of the *Leishmania* genome

Conclusions

Thus, due to occurrence of all these possible events:
although cells in the population have a common core genome…

→ Many individual cells will differ from the rest of the population by carrying one or more distinct amplicons.

→ Concept of **Sub-population** : Upon selection with either drugs or culture conditions, a subpopulation can emerge where the amplicon copy number per cell increases.

-> This clone of cells can then expand to dominate the population !!!

(Ouellette et al., 2014)

Complexities of the *Leishmania* genome

Conclusions

- *Leishmania* might be using **adaptive gene amplification at a genome-wide scale** as one strategy to adapt to a changing environment.
- Thinking in terms of populations, rather than individuals, is primary: the genetic diversity existing in natural populations is a key factor in evolution.
- Genetic events = DNA: best component to understand and follow the role of predictive modifications

To sequence or not to sequence!?!