NGS and non-model species
(what they don't teach you at school)

Agnieszka Kloch

Zakład Ekologii
Wydział Biologii
Uniwersytet Warszawski

Why non-model species?

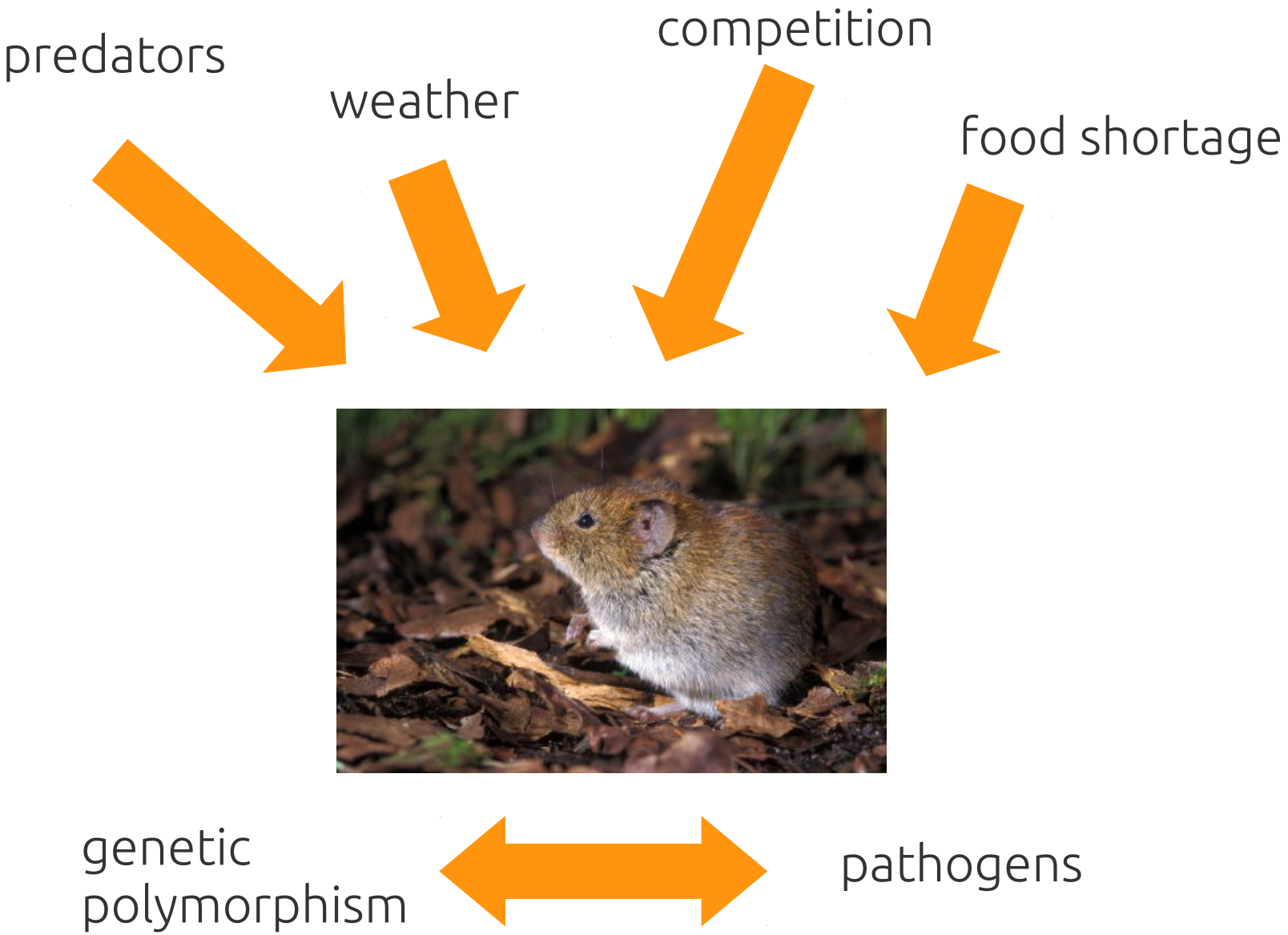
- $2*10^6 – 10^{12}$ species
- ~1000 animals sequenced
- the world outsite medical applications doesn't use model species

susceptible

resistant

SNPs
expression levels

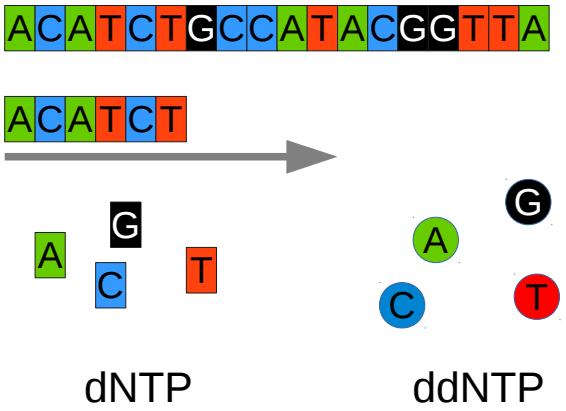
„Wild immunology”

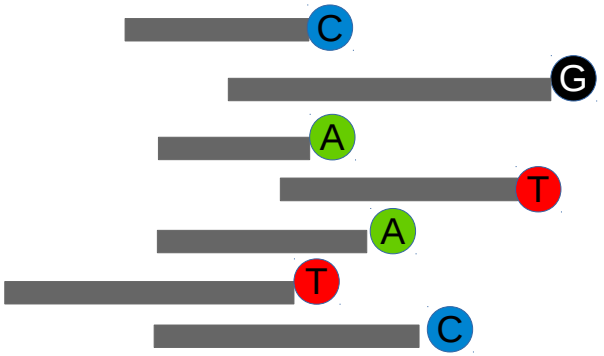


NGS technology

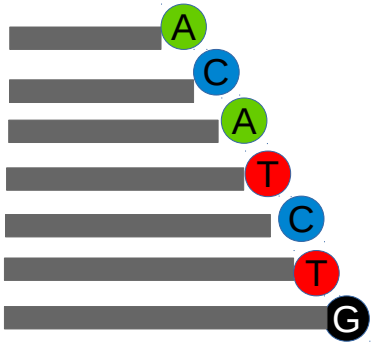
Sanger sequencing

DNA synthesis

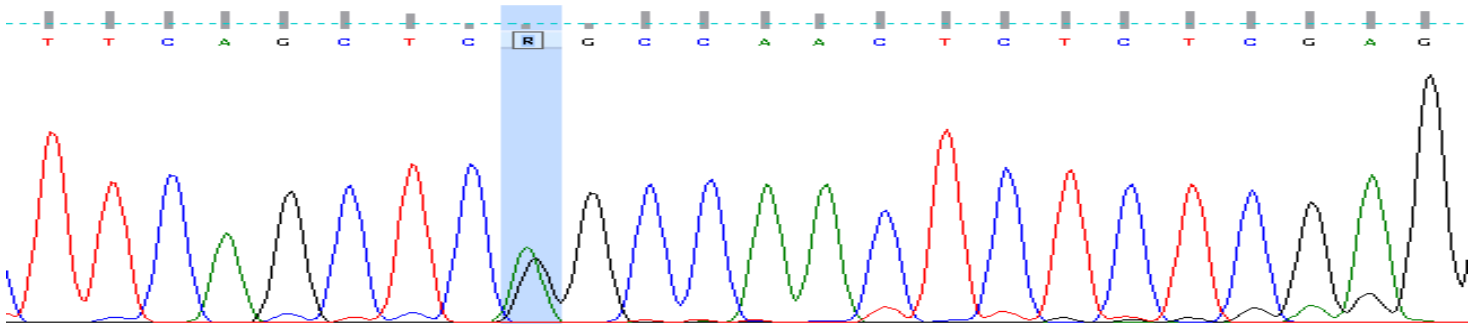




Electrophoresis

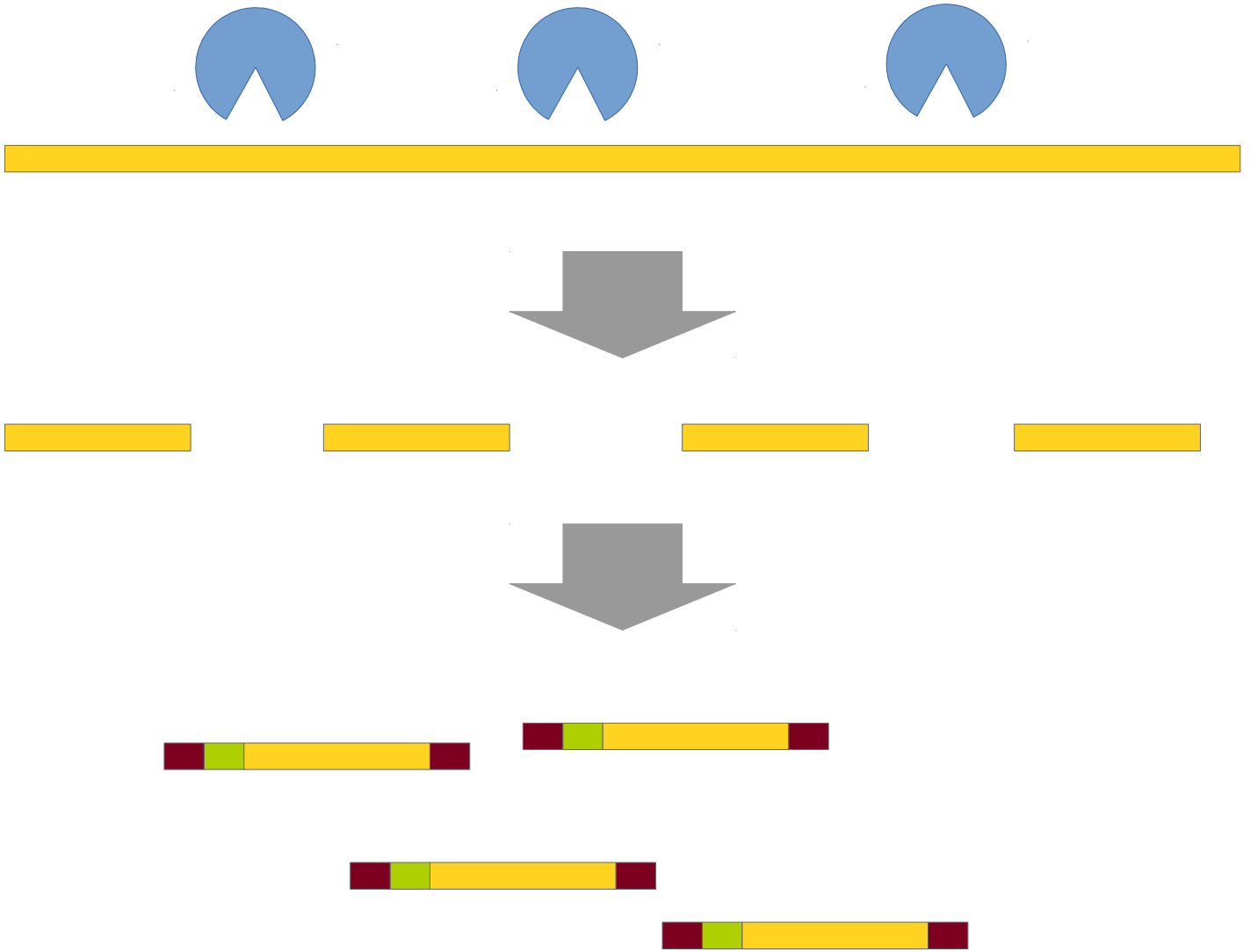


Read



NGS technology

Library preparation

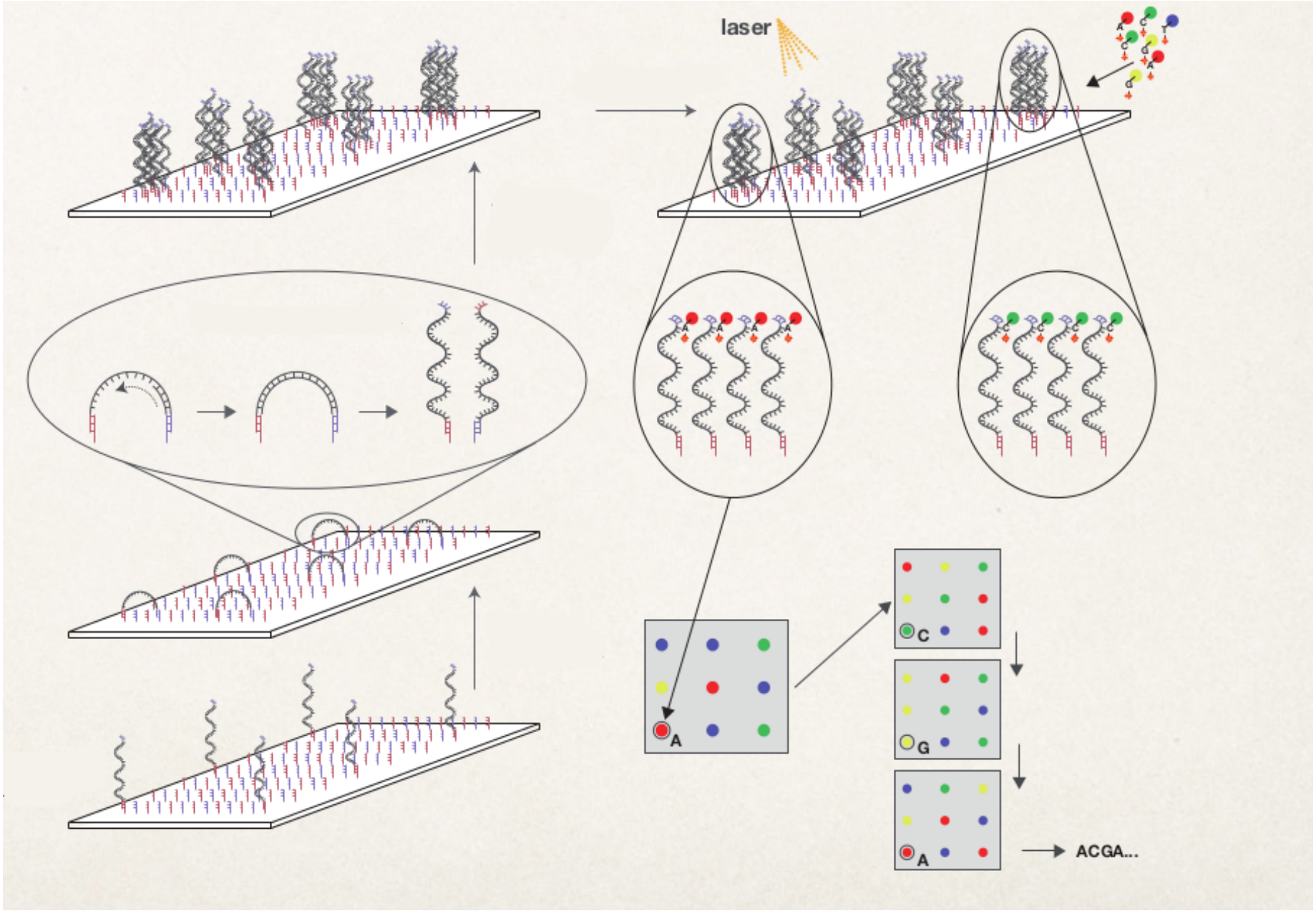


1. Cutting (fragmentation)

2. Adding adapters and barcodes

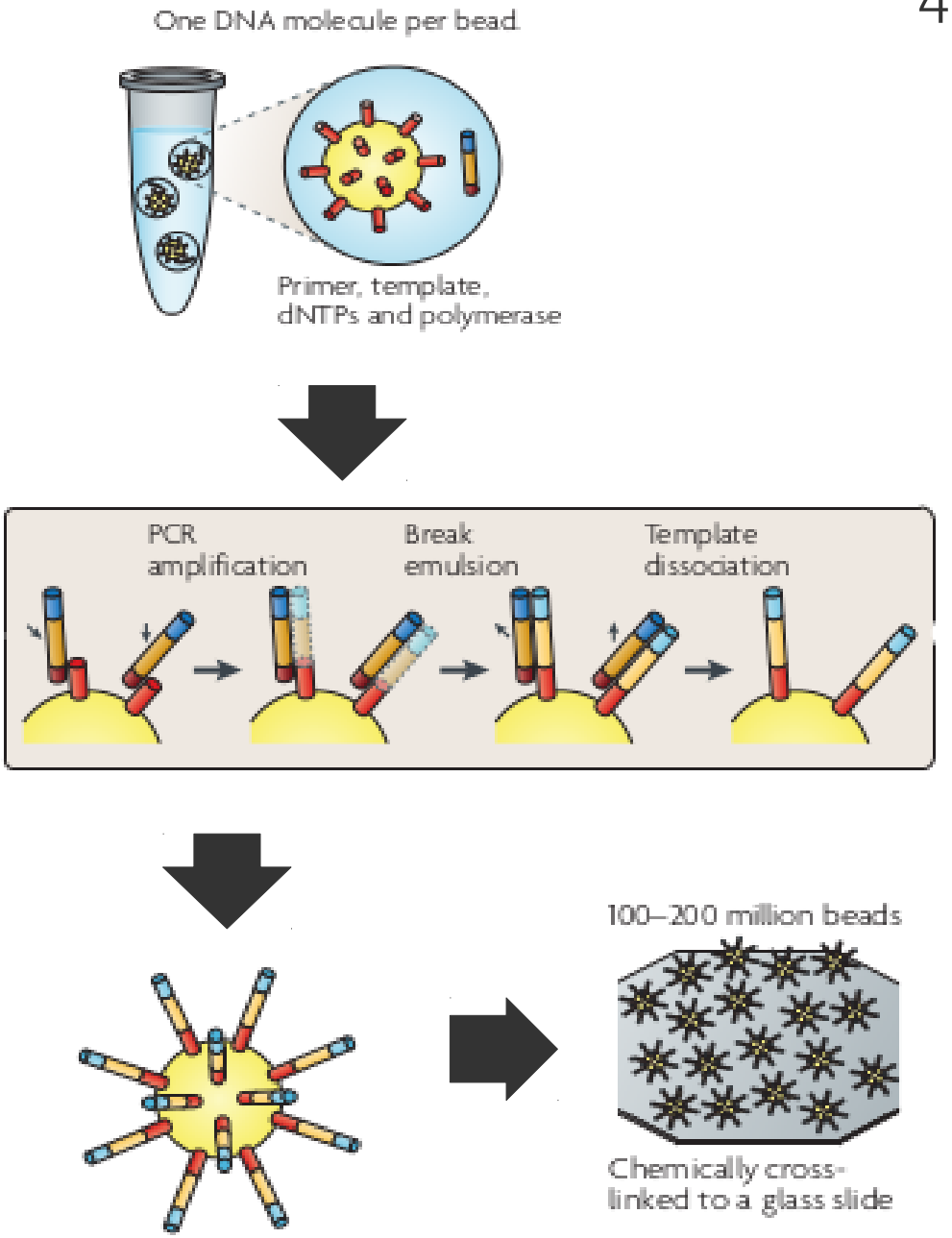
NGS technology

Illumina



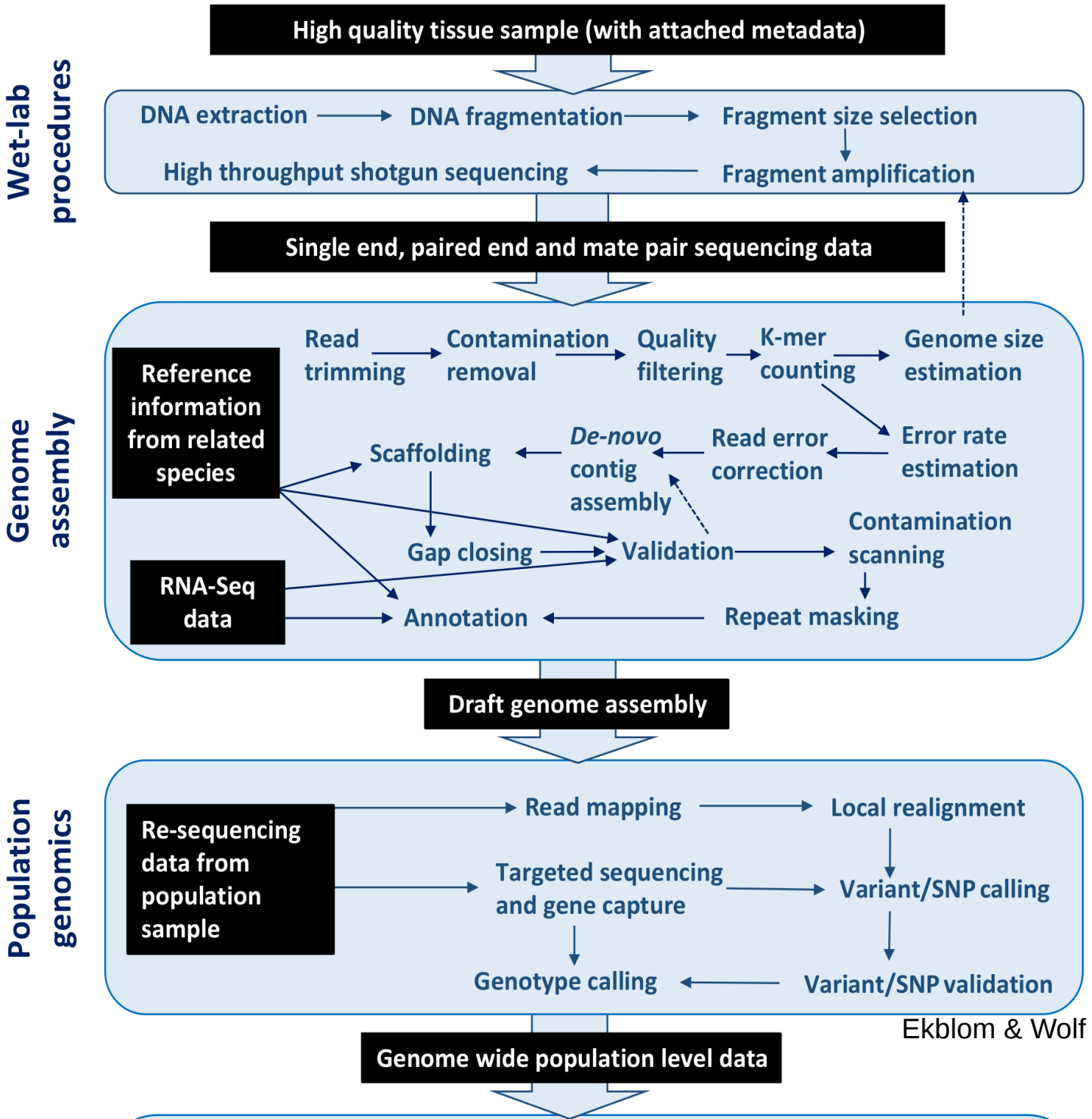
NGS technology

454 Roche



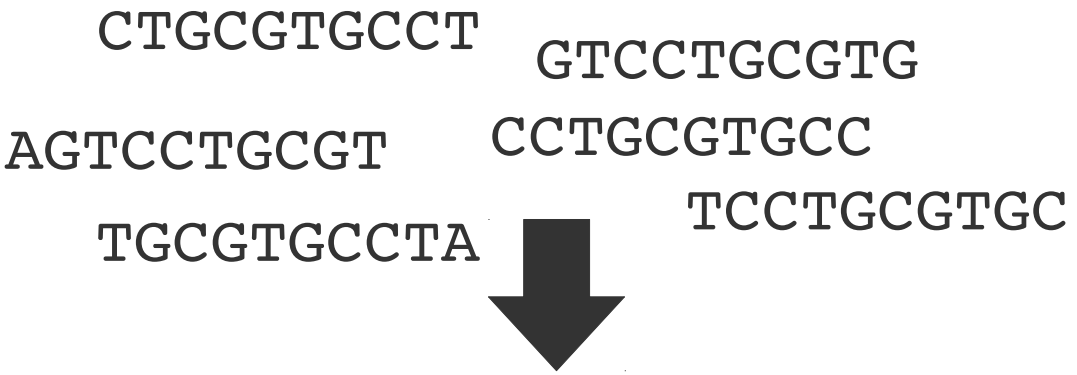
Metzker, Nat Rev Genet 11, 2010

Pipeline



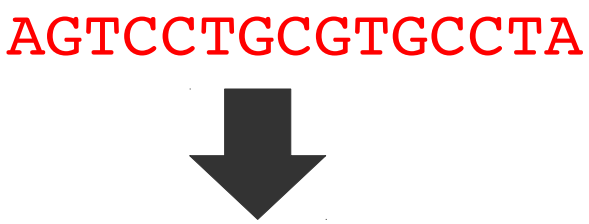
Ekblom & Wolf 2014

Assembly



reads

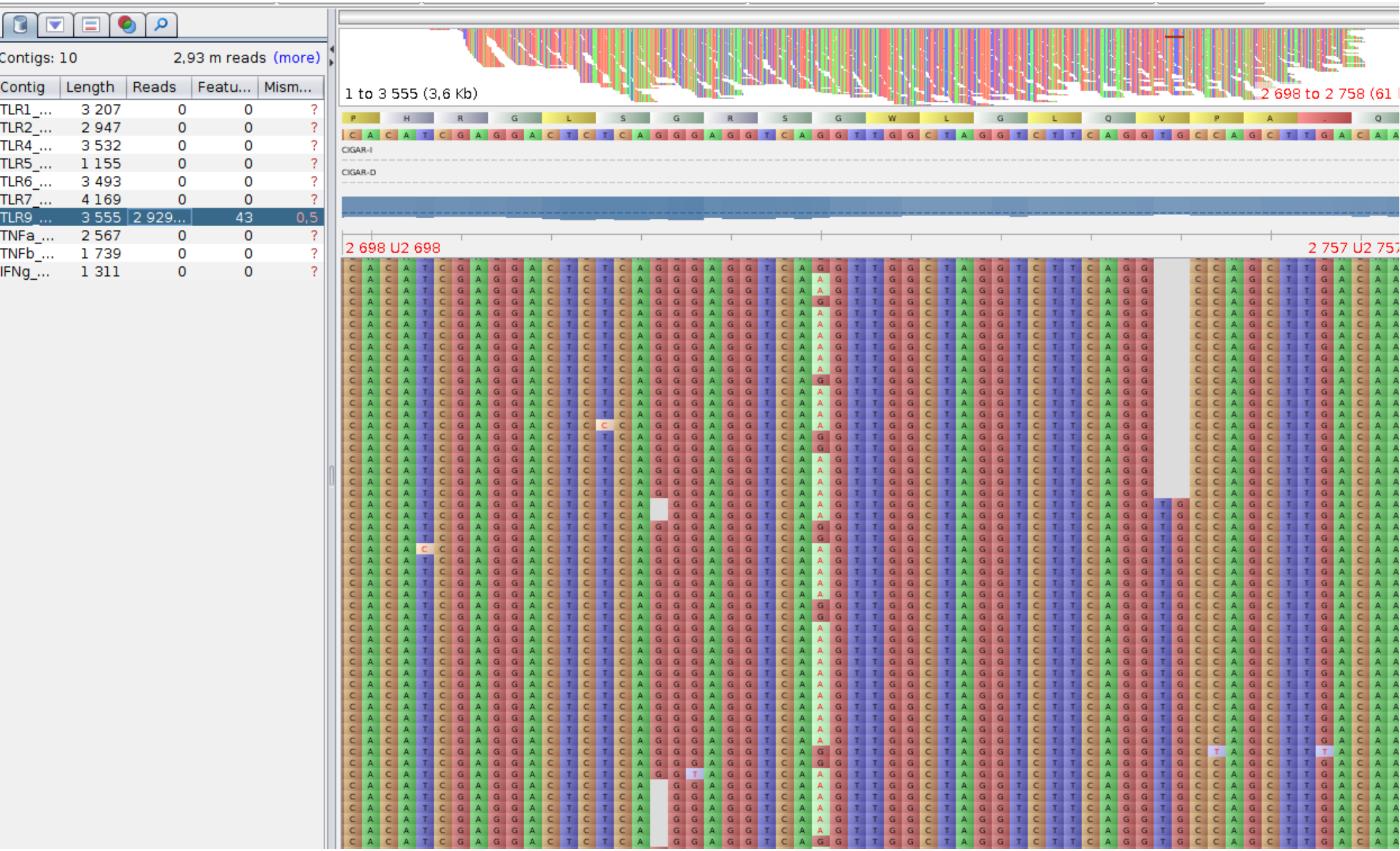
k-mers



contigi

scaffolds

SNPs


Contigs: 10
2,93 m reads (more)
Contig Length Reads Featu... Mism...
TLR1_... 3 207 0 0 ?
TLR2_... 2 947 0 0 ?
TLR4_... 3 532 0 0 ?
TLR5_... 1 155 0 0 ?
TLR6_... 3 493 0 0 ?
TLR7_... 4 169 0 0 ?
TLR9_... 3 555 2 929... 43 0,5
TNFa_... 2 567 0 0 ?
TNFb_... 1 739 0 0 ?
IFNg_... 1 311 0 0 ?
1 to 3 555 (3,6 Kb)
2 698 to 2 758 (61
CIGAR-I
CIGAR-D
2 698 U2 698
2 757 U2 757

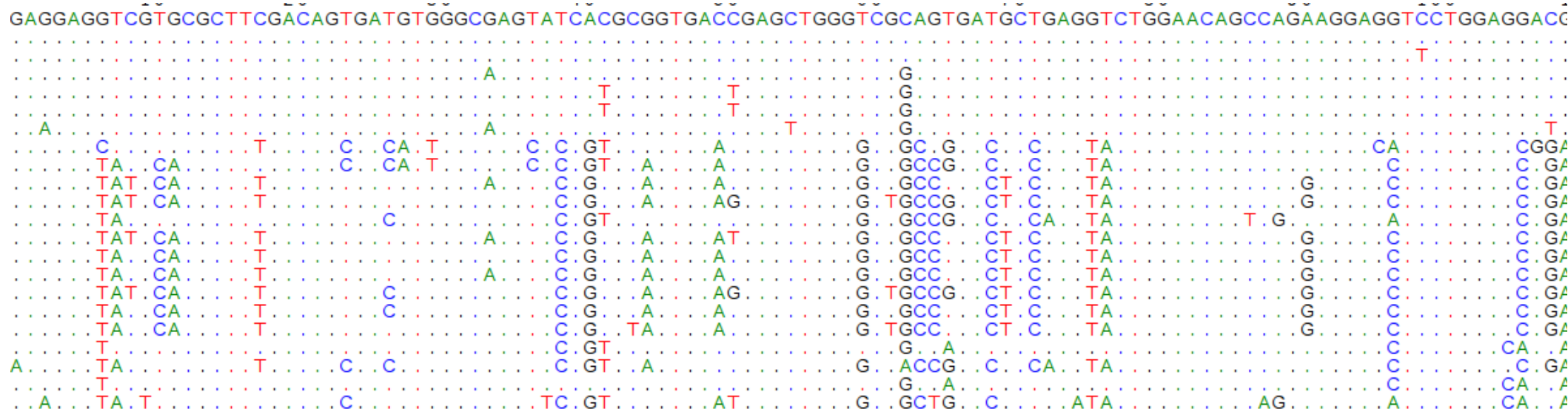
Errors!

Sources of errors:

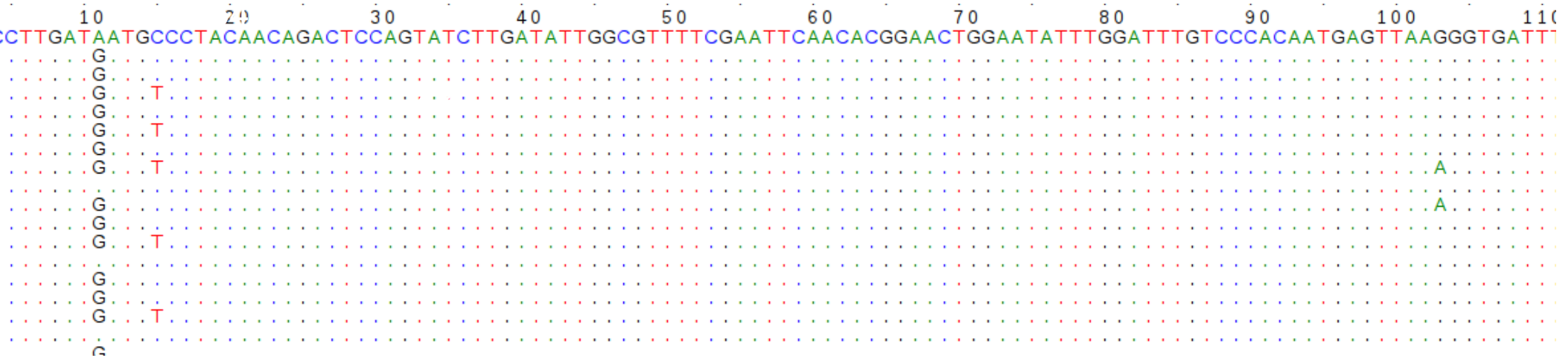
- PCR: polymerase (10^{-5} – 10^{-7})
- sequencing
- assembling

Variability in polymorphism

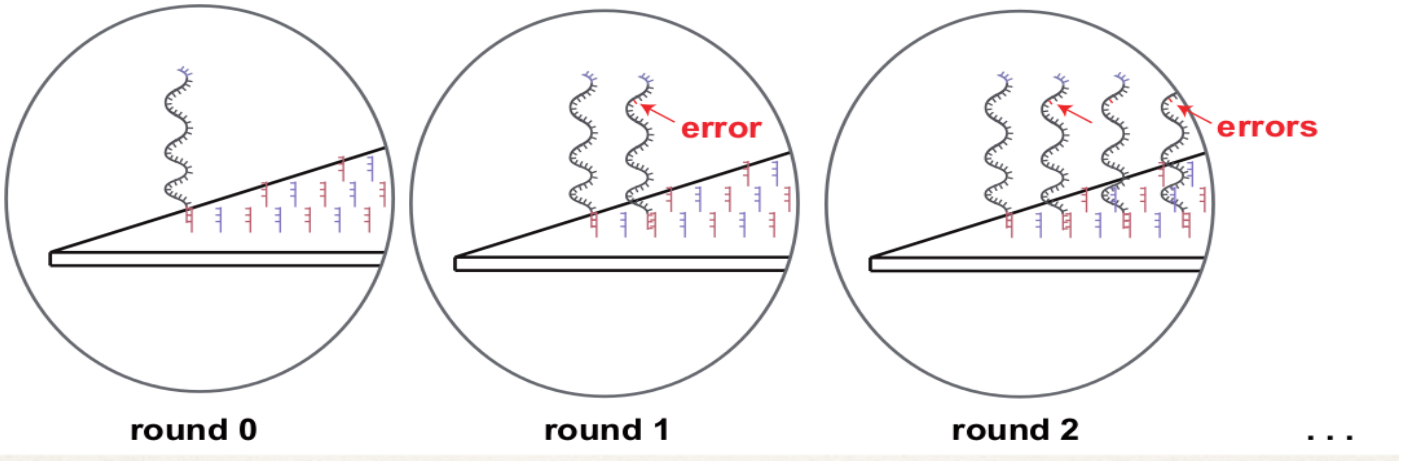
MHC



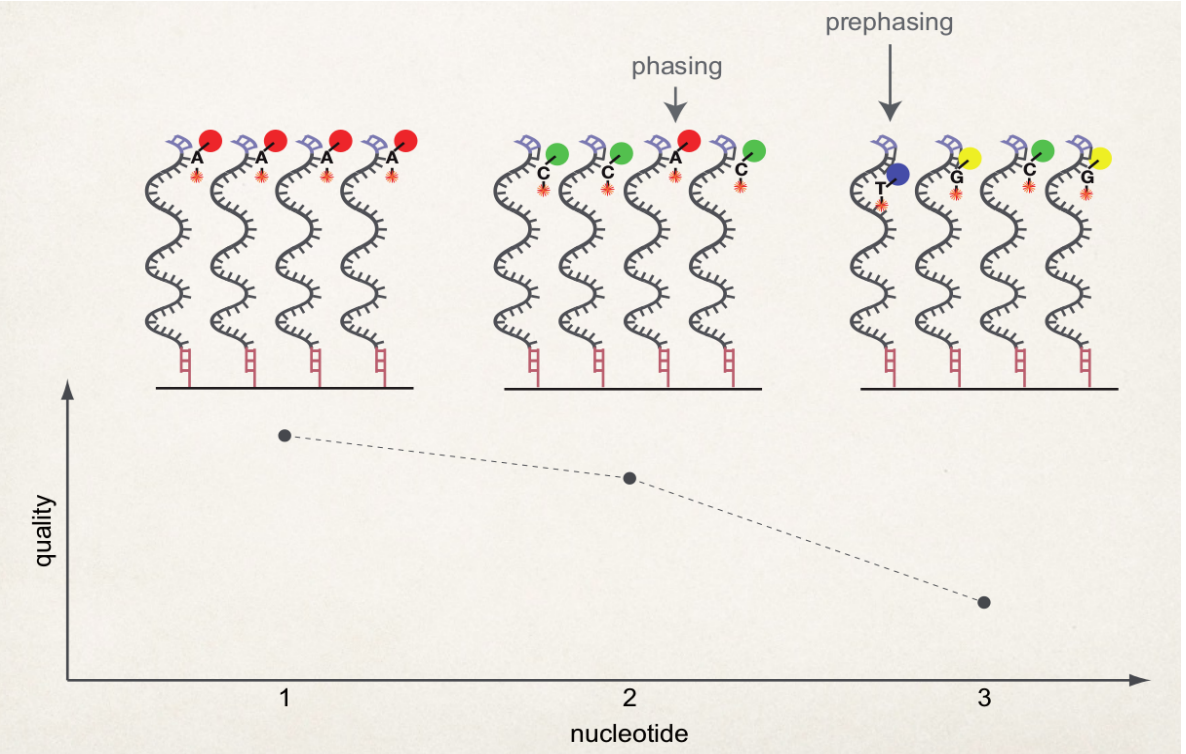
TLR



Sequencing errors



„out of phase" amplification



slide by David Langenberger

Contaminants

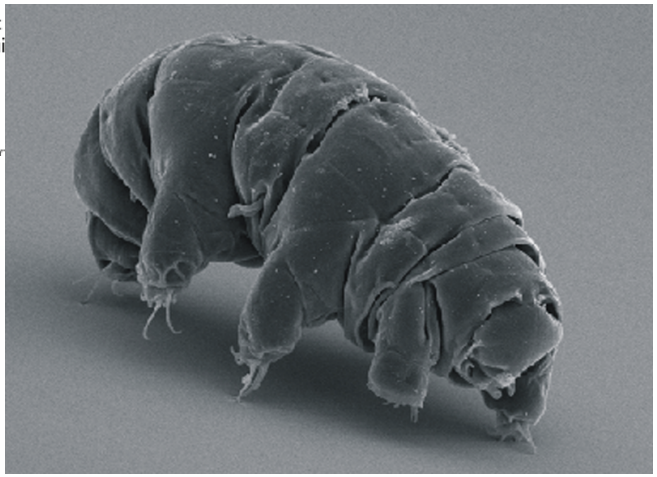
PNAS

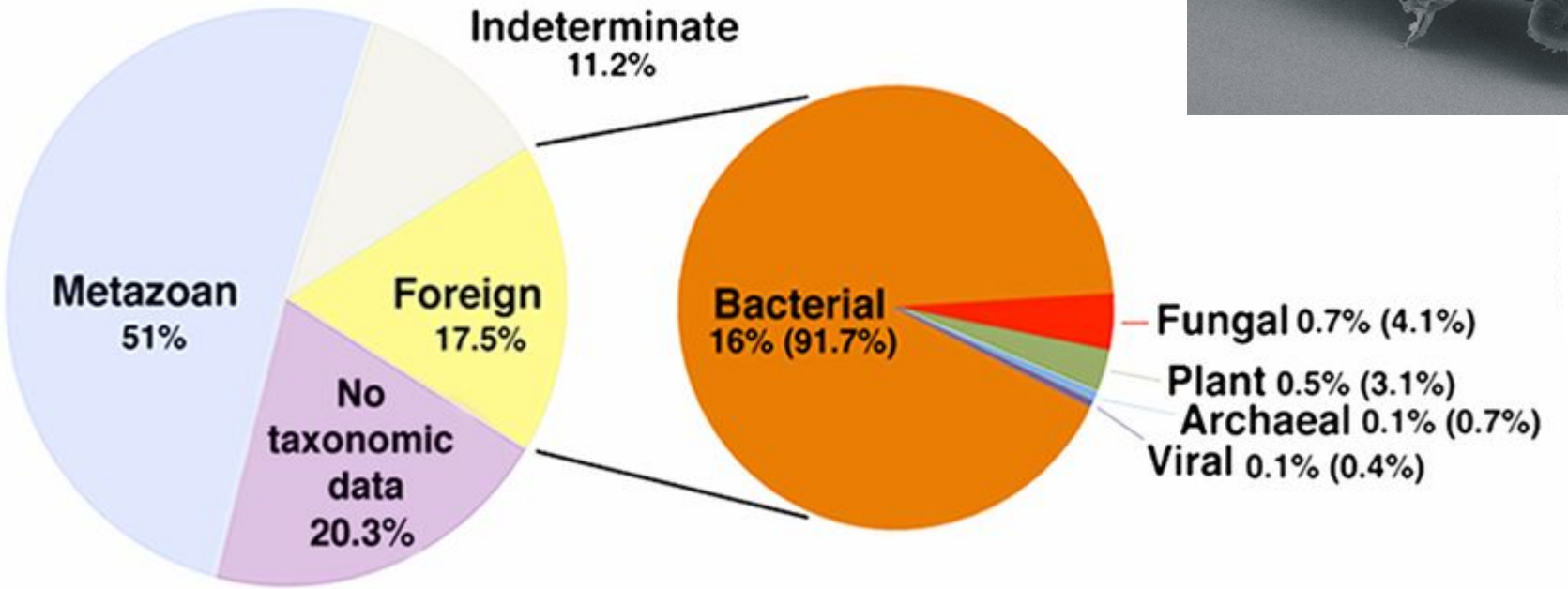
Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade

Thomas C. Boothby[a,1], Jennifer R. Tenlen[a,2], Frank W. Smith[a], Jeremy R. Wang[a,b], Kiera A. Patanella[a], Erin Osborne Nishimura[a], Sophia C. Tintori[a], Qing Li[c], Corbin D. Jones[a], Mark Yandell[c], David N. Messina[d], Jarret Glasscock[d], and Bob Goldstein[a]

[a]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [b]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [c]Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112; and [d]Cofactor Genomics, St. Loui MO 63110

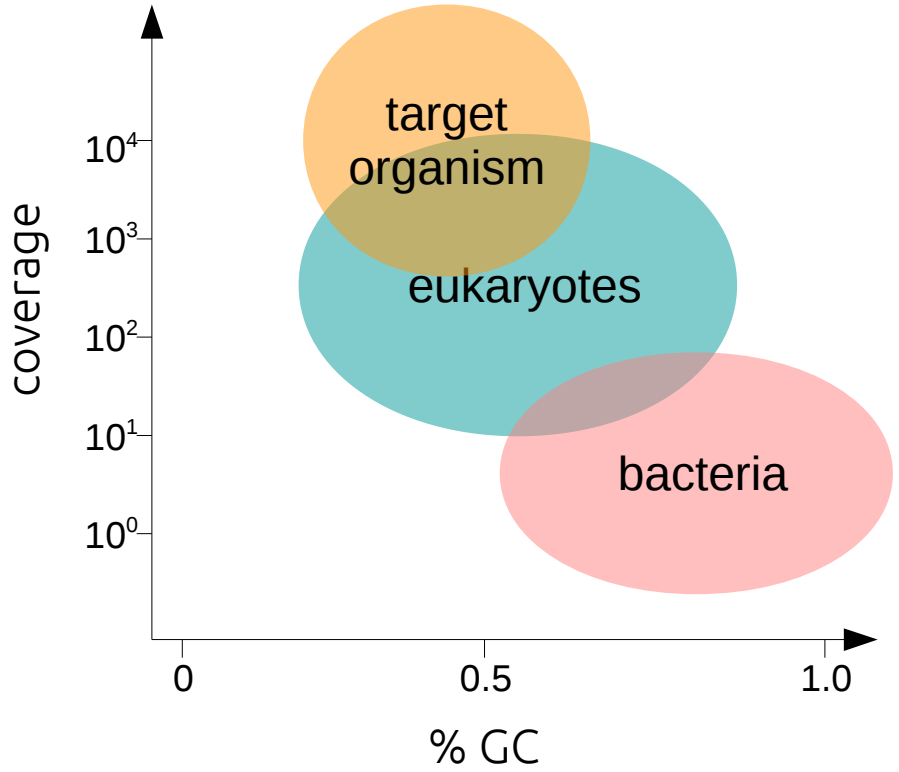
Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved September 28, 2015 (received for review May 28, 2015)

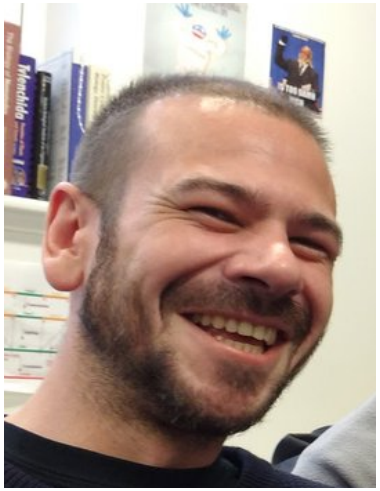




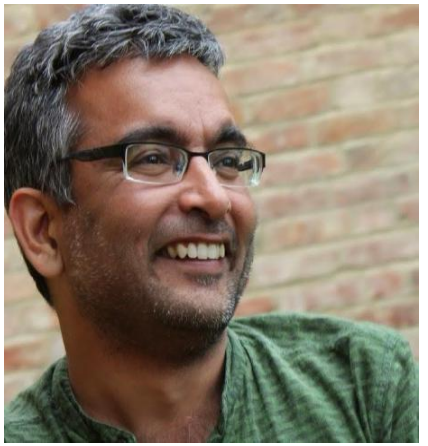
Contaminants

blobtools.readme.io





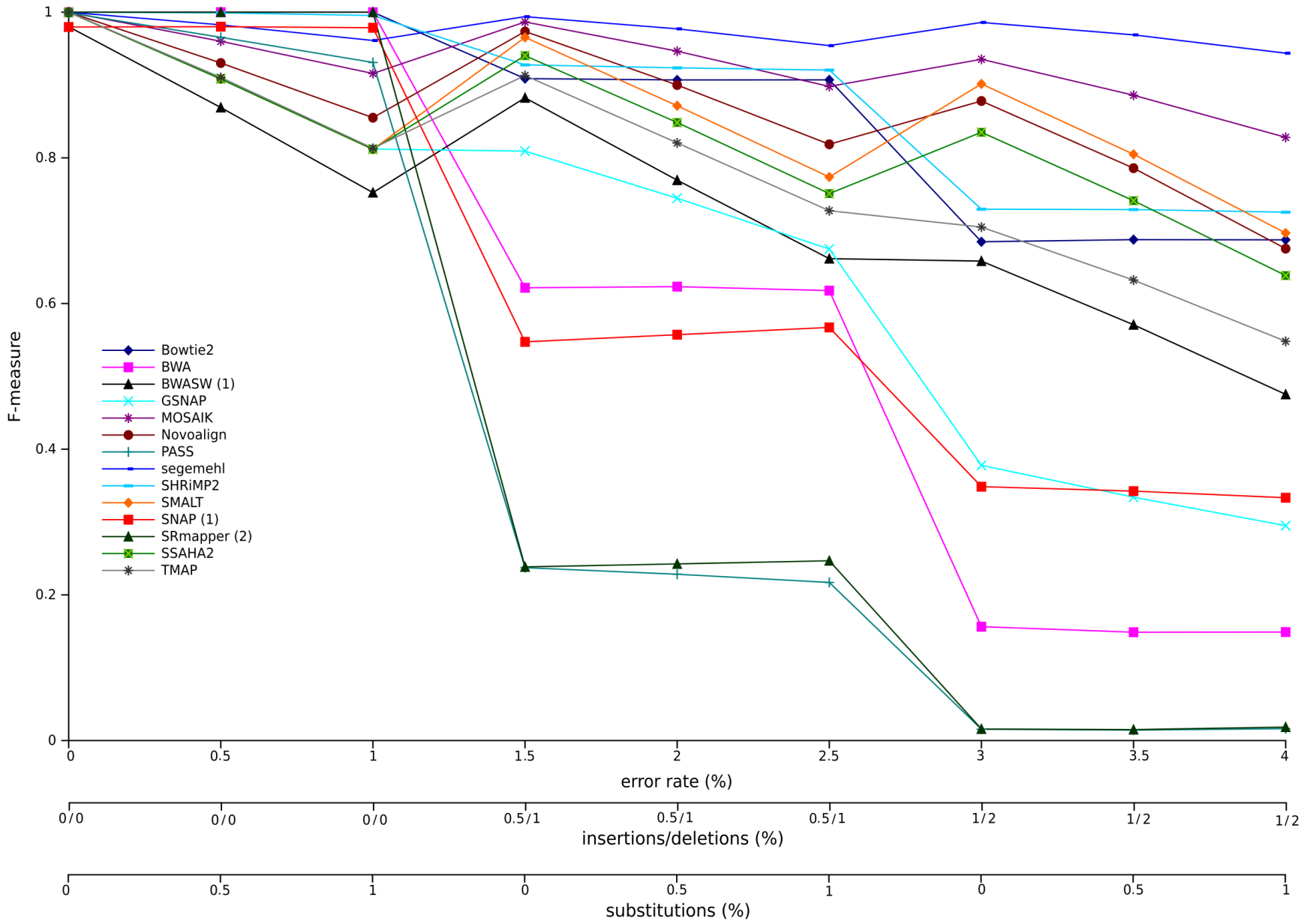
Dominik R. Laetsch



Sujai Kumar

Kumar et al. 2013. Blobology: exploring raw genome data for contaminants, symbionts, and parasites using taxon-annotated GC-coverage plots. Front Genet 4

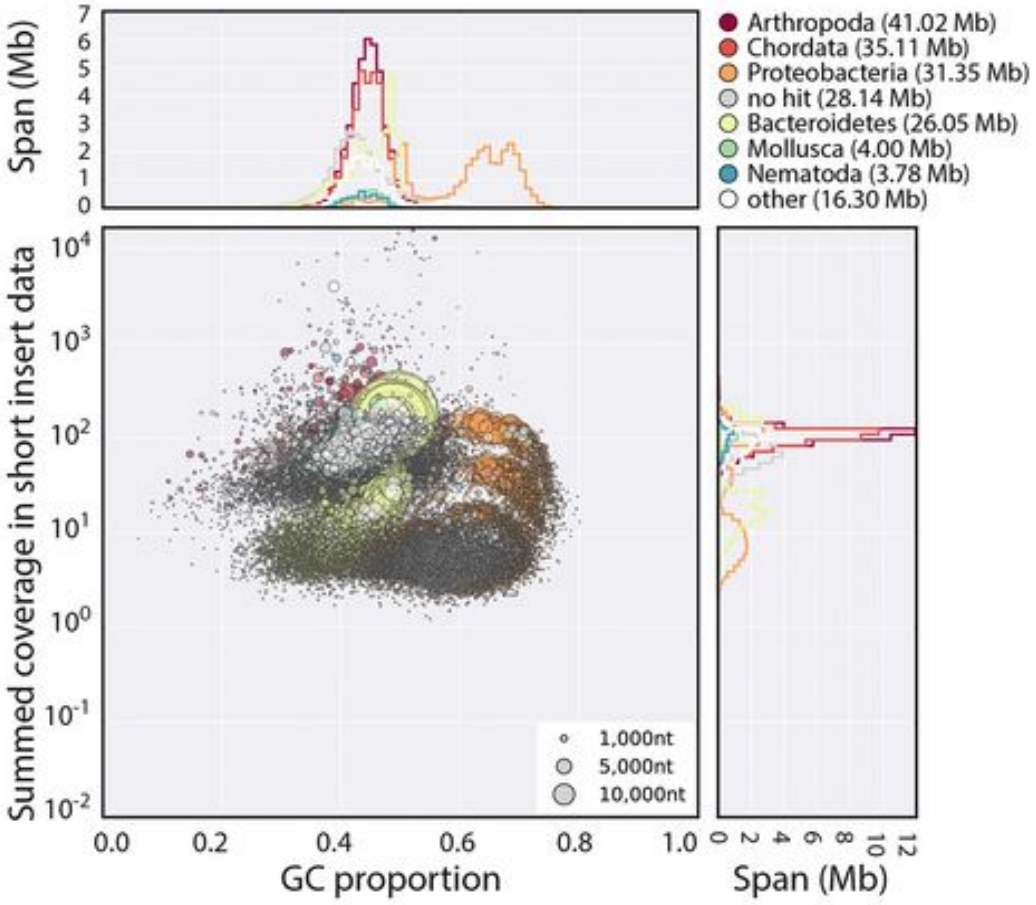
Performance of aligner software



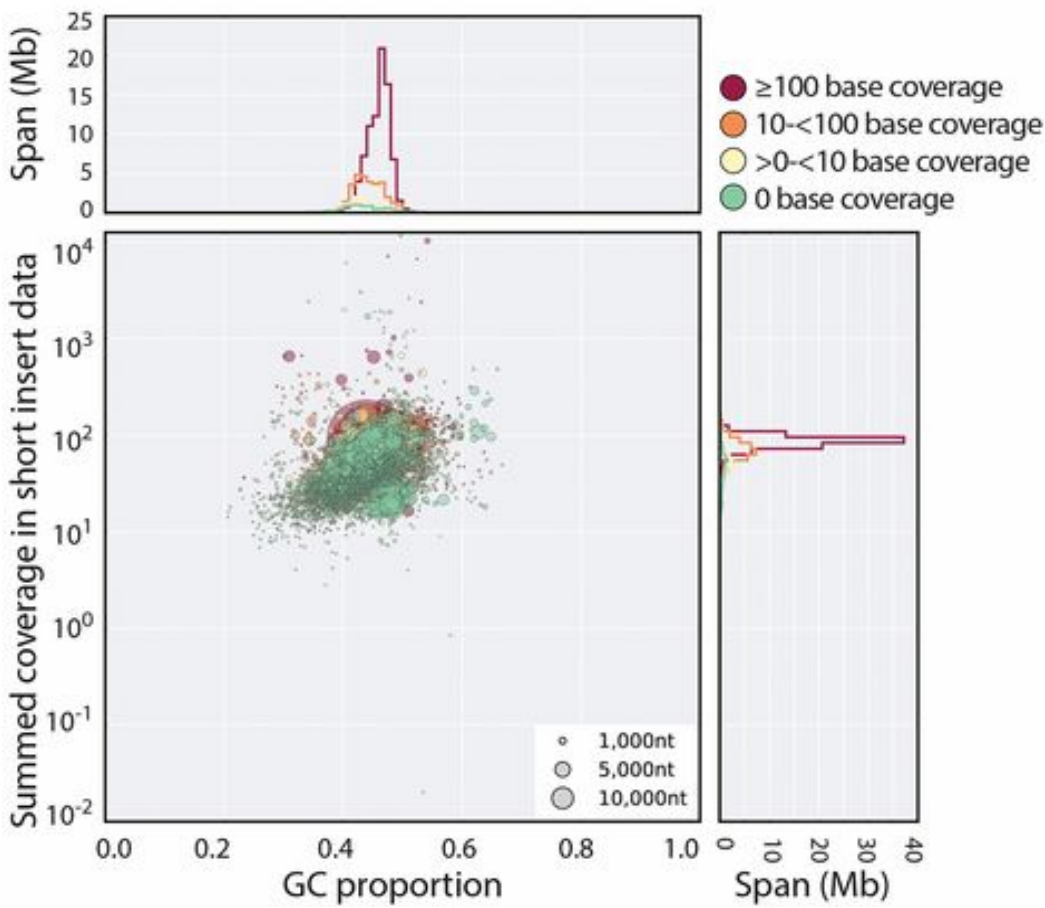
Caboche et al. BMC Genomics 2014

Contaminants

Before



After



Koutsovolous et al. PNAS 2016

Contaminants

UBSCRIBE SEARCH MENU≡ Rival Scientists Cast Doubt Upon Recent Discovery About Invincible Animals

Rival Scientists Cast Doubt Upon Recent Discovery About Invincible Animals

A recent claim that tardigrades got a sixth of their DNA from microbes is starting to unravel.

5.0k

TEXT SIZE

ED YONG | DEC 4, 2015 | SCIEN

Last Monday, a team from the U
published the first ever genome
microscopic animals with a repu

ScienceNews
MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

Subscribe | Advertise
Archive
Search Science News...

Explore

LATEST | MOST VIEWED

NEWS IN BRIEF
Physicists find signs of four-neutron nucleus
BY ANDREW GRANT FEBRUARY 08, 2016

SCIENCE TICKER
This roach-inspired robot can wiggle through tight spaces
BY SARAH SCHWARTZ FEBRUARY 08, 2016

NEWS IN BRIEF
Cancer drug's usefulness against Alzheimer's disputed
BY LAURA SANDERS FEBRUARY 08, 2016

NEWS
Don't blame winter for that bleak mood
BY BRUCE BOWER FEBRUARY 08, 2016

TELEVISION

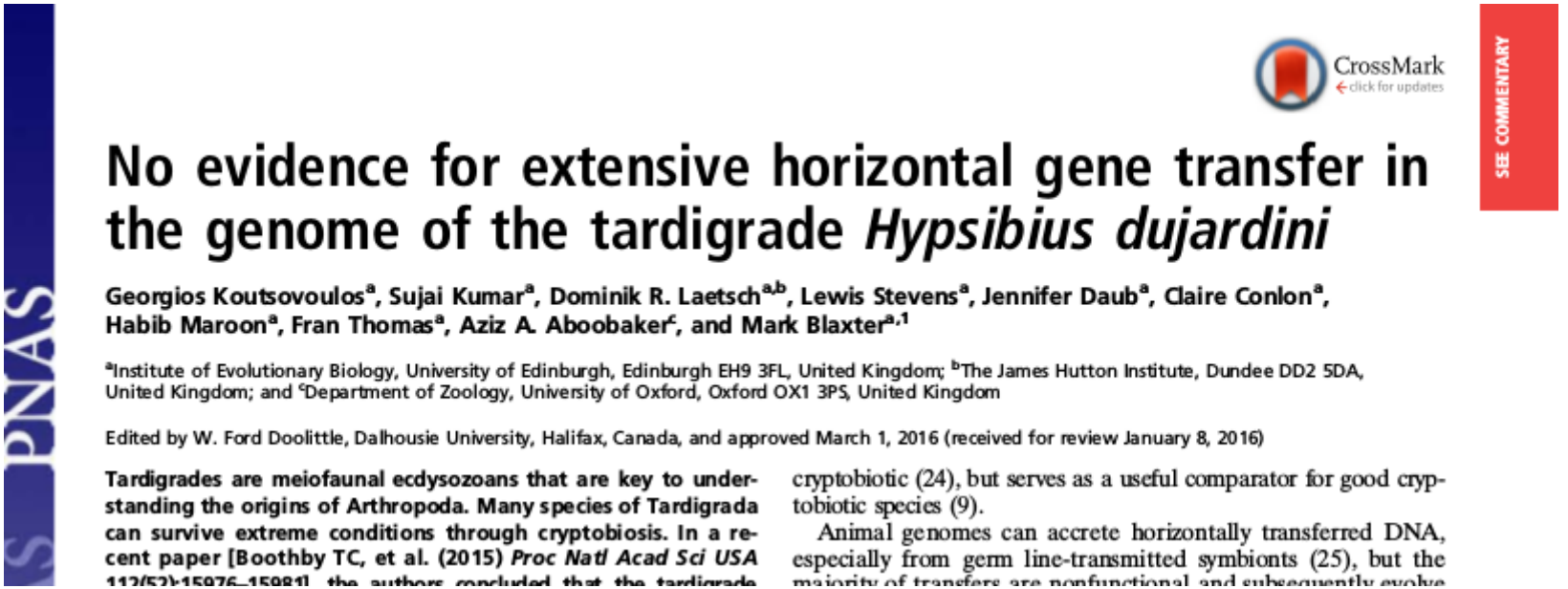
NEWS ANIMALS, GENETICS

Water bears' genetic borrowing questioned

A new analysis finds bacterial DNA in the tardigrade genome is mostly contamination

BY TINA HESMAN SAEY 3:55PM, DECEMBER 8, 2015

Contaminants



PNAS

CrossMark click for updates

SEE COMMENTARY

No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*

Georgios Koutsovoulos[a], Sujai Kumar[a], Dominik R. Laetsch[a,b], Lewis Stevens[a], Jennifer Daub[a], Claire Conlon[a], Habib Maroon[a], Fran Thomas[a], Aziz A. Aboobaker[c], and Mark Blaxter[a,1]

[a]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom; [b]The James Hutton Institute, Dundee DD2 5DA, United Kingdom; and [c]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved March 1, 2016 (received for review January 8, 2016)

Tardigrades are meiofaunal ecdysozoans that are key to understanding the origins of Arthropoda. Many species of Tardigrada can survive extreme conditions through cryptobiosis. In a recent paper [Boothby TC, et al. (2015) *Proc Natl Acad Sci USA* 112(52):15976–15981], the authors concluded that the tardigrade

cryptobiotic (24), but serves as a useful comparator for good cryptobiotic species (9).

Animal genomes can accrete horizontally transferred DNA, especially from germ line-transmitted symbionts (25), but the majority of transfers are nonfunctional and subsequently evolve

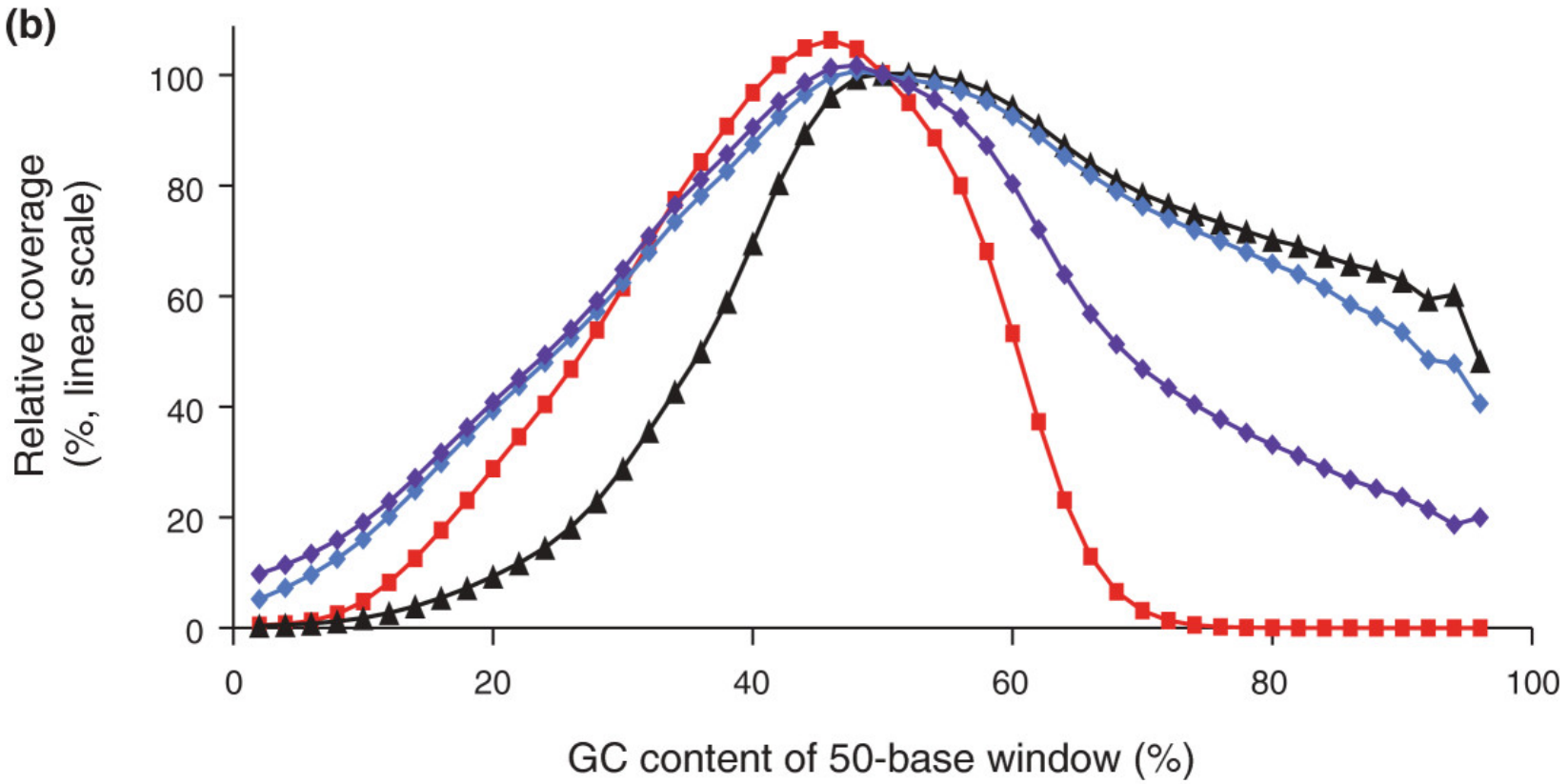
0.4% of genes likely originated through horizontal gene transfer

If there is no reference...

Potential problems:

- repetitive regions, super-variable regions
- coverage not equal in all regions
- various ploidy, often unknown
- contaminants

GC bias



Illumina reads, 400-bp fragment library prepared with:

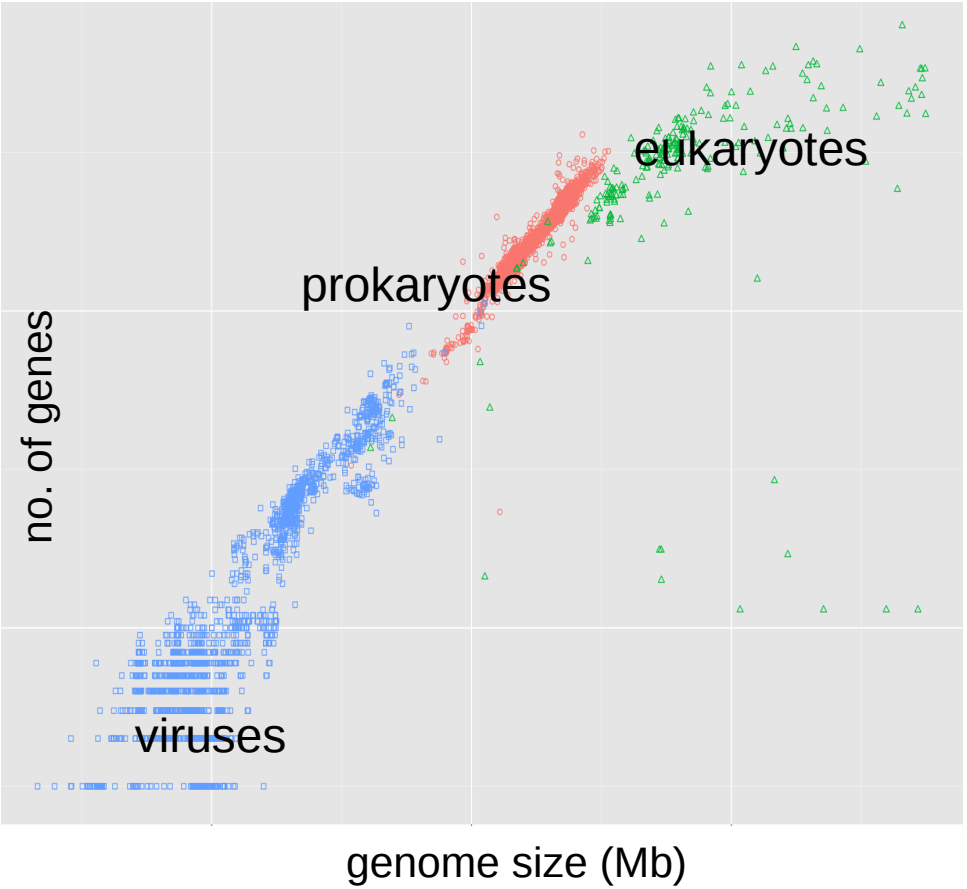
- standard PCR protocol, fast-ramping thermocycler (red squares),
- Phusion HF with long denaturation and 2M betaine (black triangles),
- AccuPrime Taq HiFi with long denaturation and primer extension at 65°C (blue diamonds) or 60°C (purple diamonds)

Aird et al., Genome Biology 2011

Repetitive regions

	human	*C. elegans*
genome (Gb)	3.2	0.1
genes	30 000	20 400
% coding	~1%	~25%

Genome size and number of genes



De-novo assembly of a nematode *H. mixtum*

	Full assembly	Only contings > 2000bp
span	926 859 580	287 242 474
N50	439	4347
longest scaff	32 946	32 946

Comparing assemblers (another nematode)

Metric	Sanger	PacBio Corrected	Velvet (Illumina)	Platanus (Illumina)	Platanus +SSPACE	Platanus +PBJelly	Platanus SSP+PB
No. contigs	33,365	20,661	48,965	18,039	10,555	14,734	10,574
Span (Mb)	552.85	468.39	570.53	597.80	648.50	687.69	728.17
N50 (Kb)	36.48	37.03	22.81	91.18	248.07	145.75	281.43
N's span (Mb)	37.92	0	26.56	22.40	63.09	4.19	26.34
CEGMA P	83.87	56.05	87.10	90.73	91.94	92.74	
CEGMA C	52.02	35.89	47.58	57.26	58.87	58.06	
Transcriptome No hits	3,625	12,201		1,768	1,768	1,743	
Transcriptome 70%	25,116	19,851		28,651	29,663	29,524	

Koutsovoulos 2015

De-novo annotation

No transcriptome for *H. mixtum*

Gene predictions from:

- similar species transcriptomes (*Rhabiditida*)
- typical features of a gene
- orthology, Core Eukaryotic Genes (CEGMA)

General summary	
genes	85 496
mRNA	85 496
CDS	214 889
3'UTR	650
5'UTR	3 240
exon	215 224

Core Eukaryotic Genes (248 ultra-conserved CEG)		
	complete	partial
% identified	23.7	54.8
no. of orthologs per CEG	1.53	1.9
% multiple orthologs	40.6%	58.8

De-novo assembling and annotation

Solution

- combine Illumina with PacBio
- compare sequencing data with a transcriptome
- use various assemblers

→ expensive and mundane

And there still will be errors...

Data analysis: rare alleles

`glm (infection_level ~ allele1 + allele2 + … )`

- How many alleles one can fit to a model?
- Hundreds of SNPs, dozens of individuals...

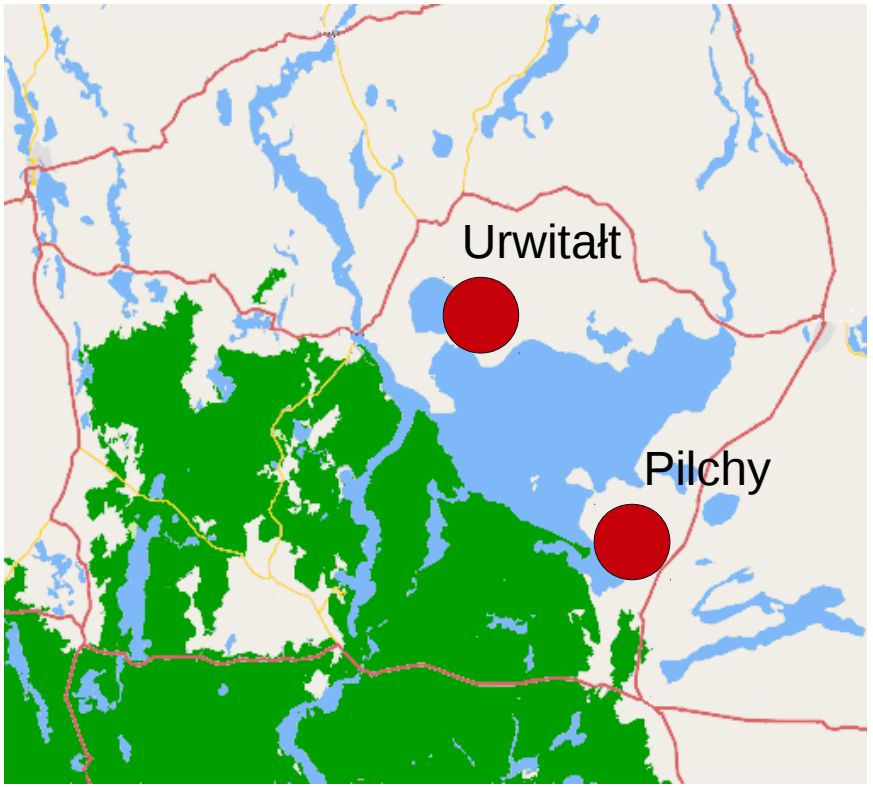
Solutions

- GWAS, MAF >0.05
- DAPC (`adegenet, ade4`)
  discriminant analysis of principal componens, Jombard et al. 2010

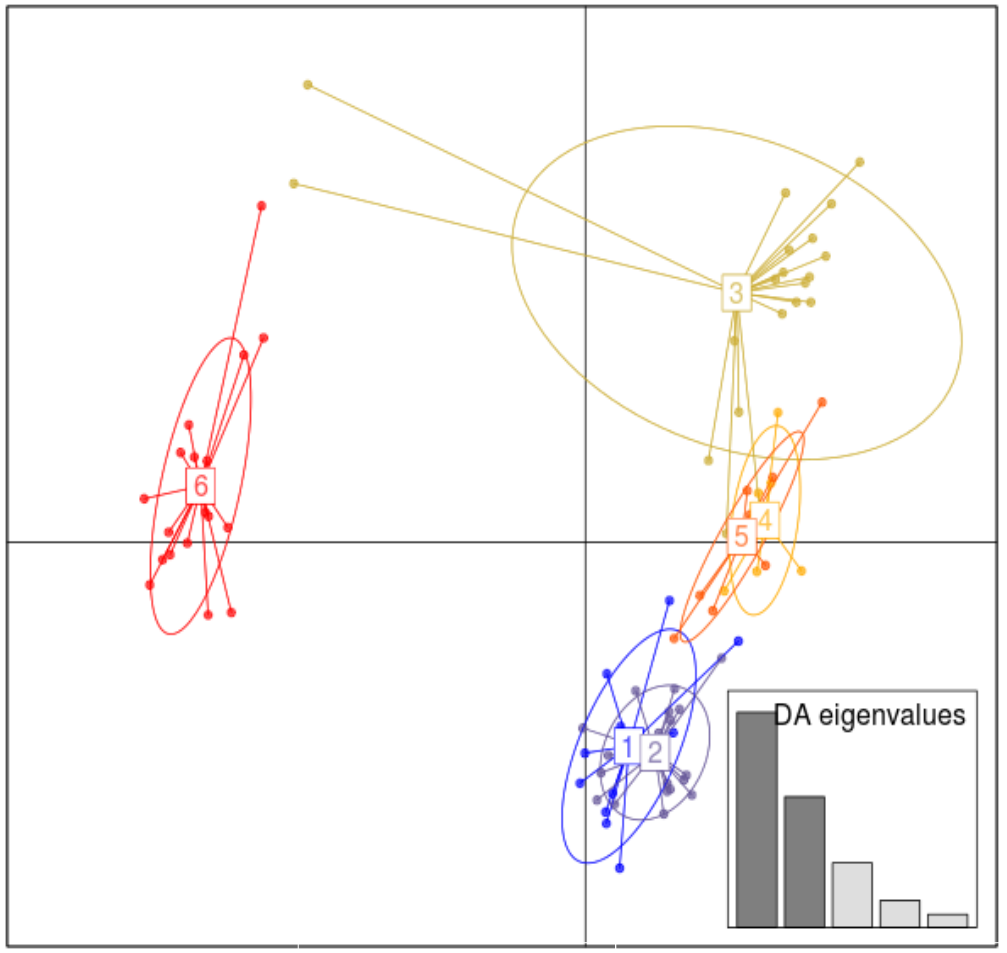
My research

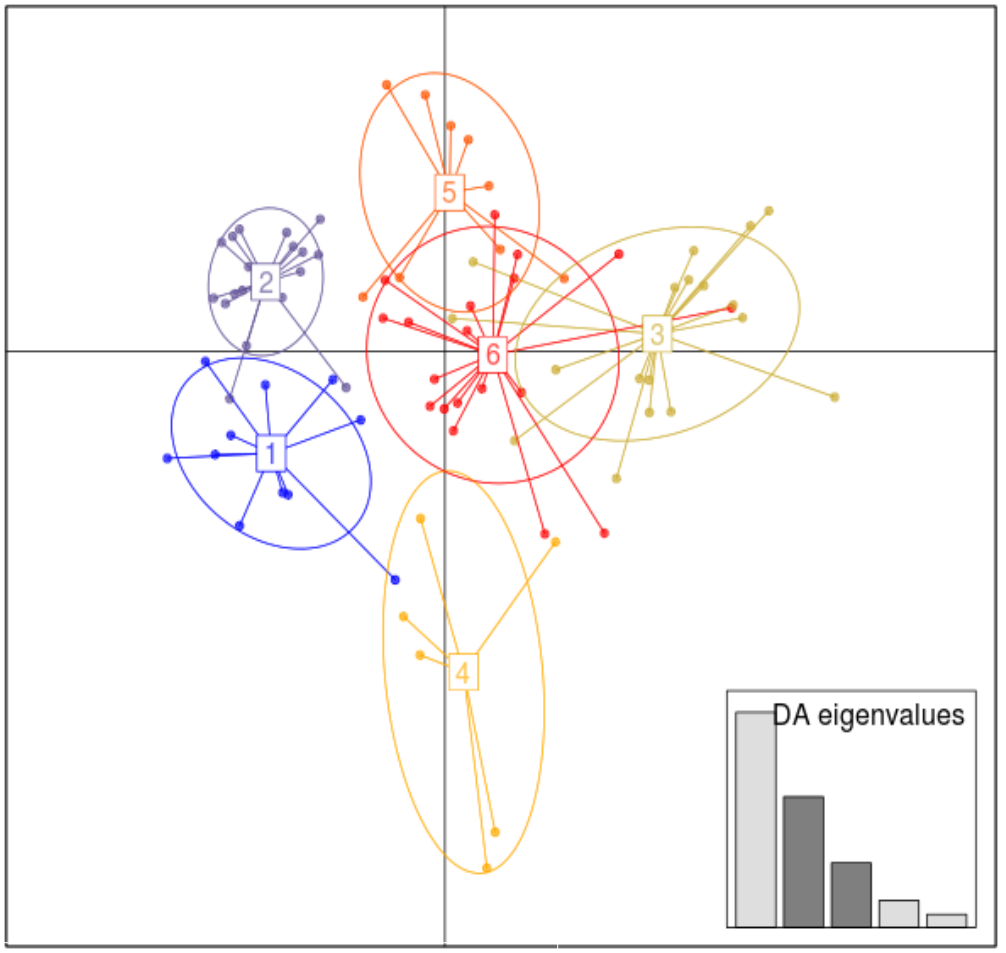
TLR genes
toll-like receptors, innate immunity

	recognized motif	primers
TLR 1	bacterial lipopeptides and lipoproteins	✔
TLR 2	bacterial lipopeptides and lipoproteins	✔
TLR 3	bacterial 23S rRNA	✘
TLR 4	bacterial lipopolysaccharides	✔
TLR 5	bacterial flagellins	✔
TLR 6	bacterial and fungial lipopeptides	✔
TLR 7	viral ssRNA	✔
TLR 8	G-rich oligonucleotides	✘
TLR 9	unmethylated CpG sequences	✔
TLR 10	not known	✘
TLR 11	not known	✘
TLR 12	not known	✘



DAPC



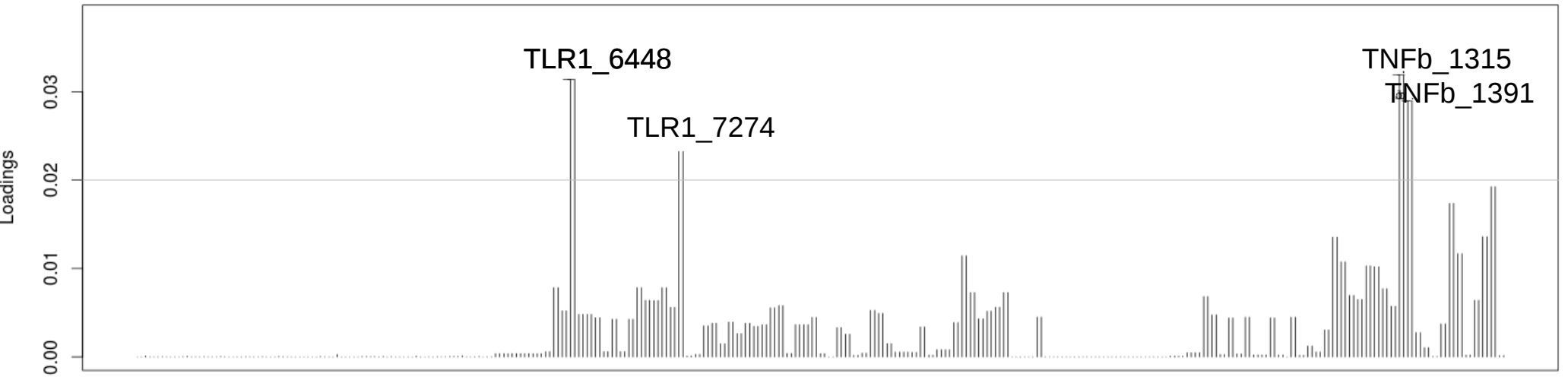


cluster

	I	II	III	IV	V	VI
Urwitałt	1	4	20	5	9	4
Pilchy	1	15	0	1	0	15

DAPC: loadings

Infection with blood parasite *Babesia*



SNPs

Life is complex

ecological factors

genetic factors

♀♂

TLR, MHC

kinship
demographic effects

A statistical riddle

data:

- 84 individuals from 2 sites
- ~200 SNPs
- ~20 parasites
- Parasite load depends on the site, host sex and hosts body mass

How to find SNPs that are associated with parasite susceptibility/resistance?

Dziękuję za uwagę

Badania finansowane z grantu NCN Opus UMO-2012/07/B/NZ8/00058

support: