Assessment of poly-guanine tracts content in *Mus musculus* genome

Daria Sergeeva
Advisors: P.Dobrynin, A. Komissarov
Dobzhansky Center for Genome Bioinformatics

Task:

To find out, whether or not poly-guanine tracts in *Mus musculus* genome are sequencing artifacts

Why they may be artifacts?

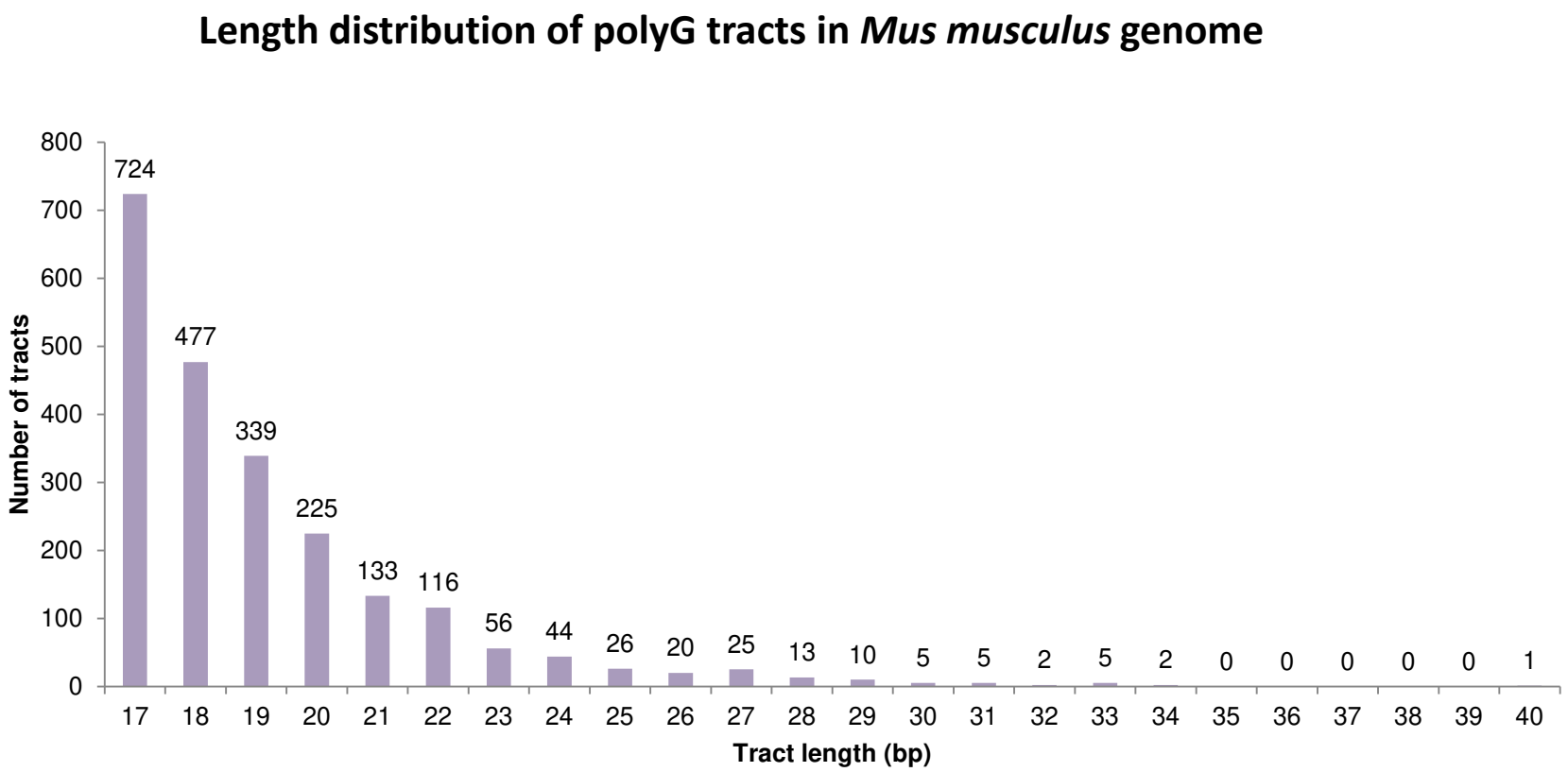
- high melting temperature
- capacity to form secondary structures, which can restrict replication and transcription
- coding of poly-Gly peptides

Pipeline

NCBI	• Download assembled chromosomes of *Mus musculus* (GRCm38.p1)
Python	• Search of polyG tracts (len>=17bp) with 200bp flanks (output – BED file)
Bowtie2	• Alignment of Illumina reads (SRR067652) to reference genome (GRCm38.p1)
SAMTOOLS	• Convert sam to bam
BEDtools	• Intersect BED file with bam file
SAMTOOLS	• Coverage depth
Python	• Convert BED to multifasta
NCBI blast	• Blast polyG tracts

Statistics

Maximal length – 40 bp (chrX)
Number of tracts - 2229
Summary length - 42665 bp



Distribution of polyG tracts on *Mus musculus* chromosomes

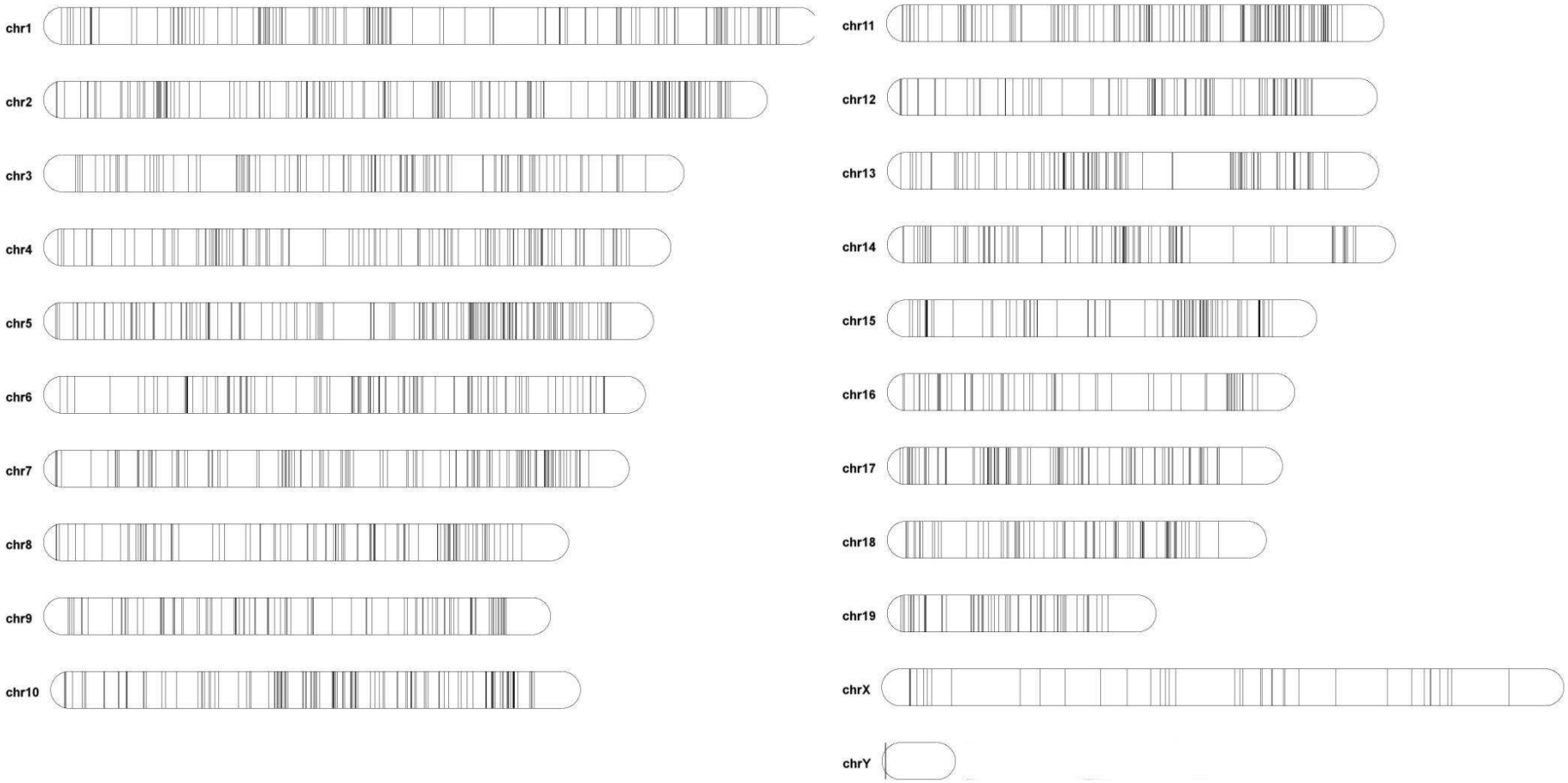


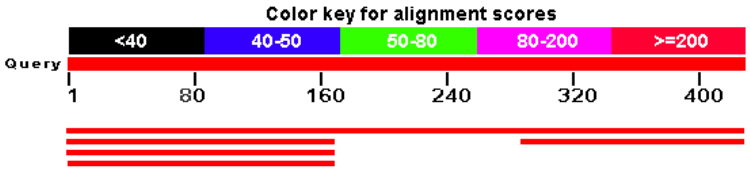
Figure provided by A. Komissarov

Results

• No one polyG tract was intersected with Illumina reads (zero coverage)

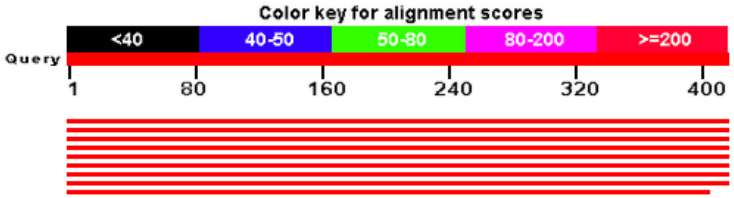
• Huge number of tracts were presented only in BAC– bacterial artificial chromosomes, which were used for genome assembly

Blast results:



	Description	Max score	Total score	Query cover	E value	Ident	Accession
→	Mus musculus 10 BAC RP24-461P19 (Roswell Park Cancer Institute (C57BL/6J Male) Mouse BAC Library) complete sequence	795	795	100%	0.0	100%	AC155943.9
	Mus musculus pleckstrin homology domain containing, family G (with RhoGef domain) member 1 (Plekhg1), transcript variant 1, mRNA	309	309	39%	1e-80	99%	NM_001159942.1
	Mus musculus adult male diencephalon cDNA, RIKEN full-length enriched library, clone:9330187D08 product:pleckstrin homology domain cor	309	309	39%	1e-80	99%	AK034403.1
	Mus musculus adult male cecum cDNA, RIKEN full-length enriched library, clone:9130218E03 product:pleckstrin homology domain containing	307	307	39%	5e-80	99%	AK033676.1
	Mus musculus 9.5 days embryo parthenogenote cDNA, RIKEN full-length enriched library, clone:B130039L12 product:pleckstrin homology dor	263	263	33%	1e-66	100%	AK045140.1

• Some polyG tracts have similarity with *Mus musculus* genes

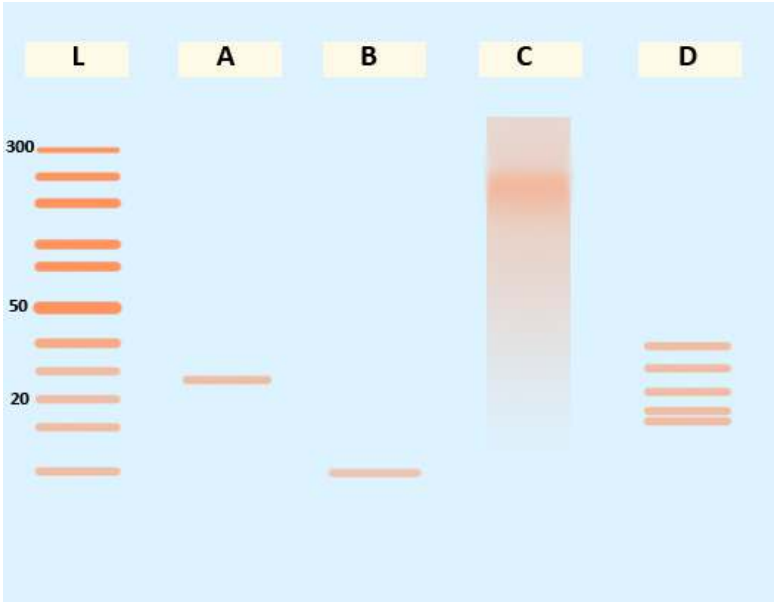


	Description	Max score	Total score	Query cover	E value	Ident	Accession
	Mus musculus targeted non-conditional, lacZ-tagged mutant allele Shc2:tm1e(KOMP)Wtsi; transgenic	773	773	100%	0.0	100%	JN952954.1
	Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele Shc2:tm1a(KOMP)Wtsi; transgenic	773	773	100%	0.0	100%	JN945047.1
→	Mus musculus C2 calcium-dependent domain containing 4C (C2cd4c), transcript variant 2, mRNA	773	773	100%	0.0	100%	NM_001168624.1
	Mus musculus C2 calcium-dependent domain containing 4C (C2cd4c), transcript variant 1, mRNA	773	773	100%	0.0	100%	NM_198614.3
	Mus musculus adult male testis cDNA, RIKEN full-length enriched library, clone:4932409I22 product:hypothetical C2 domain containing protein, full ins	773	773	100%	0.0	100%	AK029946.1

Next steps

• Pick primers for polyG tracts, which have similarities with *Mus musculus* genes

• Possible results of PCR with polyG (23bp):
L: ladder
A: polyG presented in genome
B: shorter PCR product (or no product) – flanks are in genome, but polyG is not
C: smear – flanks are from different tracts
D: many PCR products – repeated flanks or bad primers



Poly-iguanas!