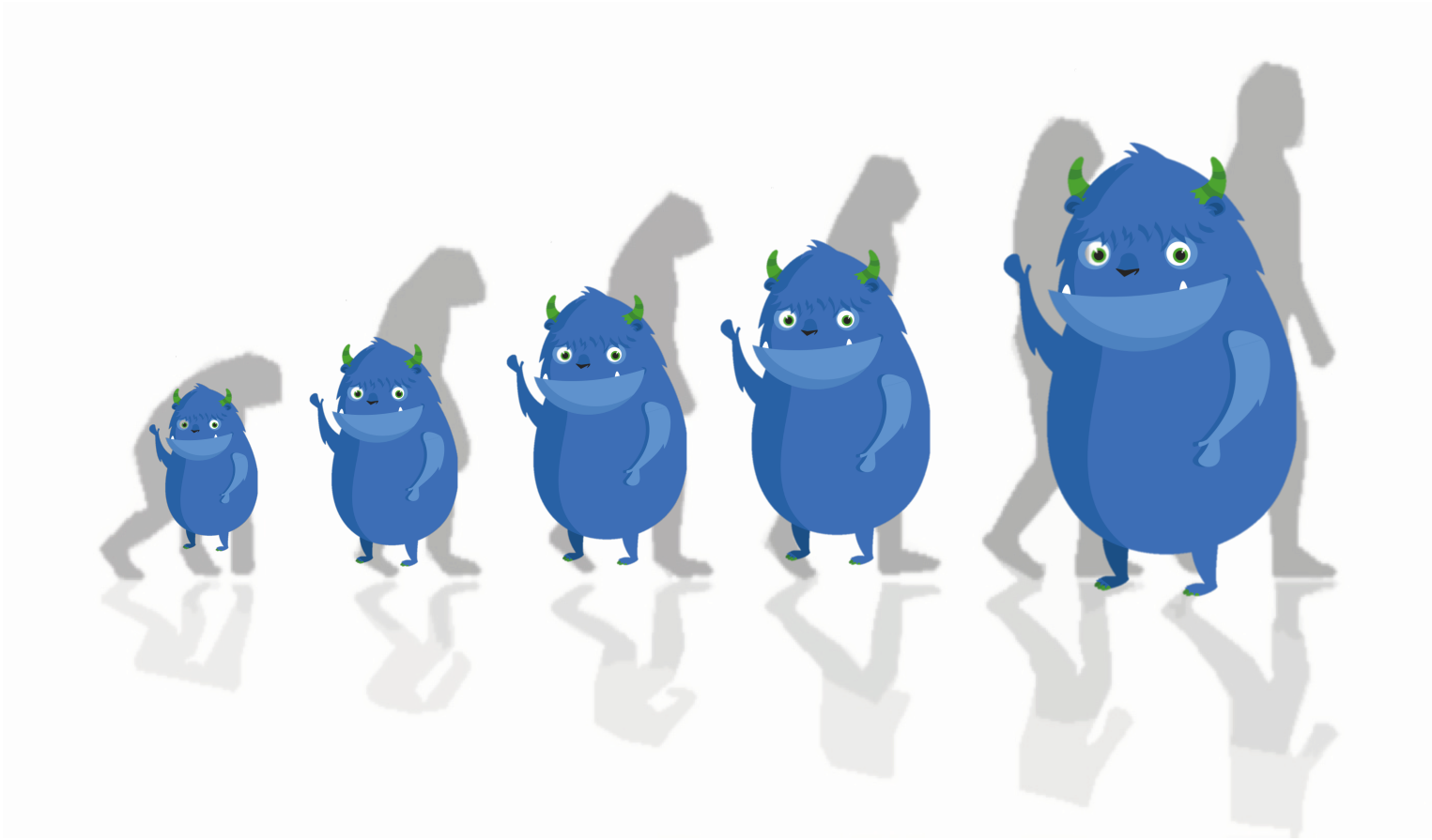
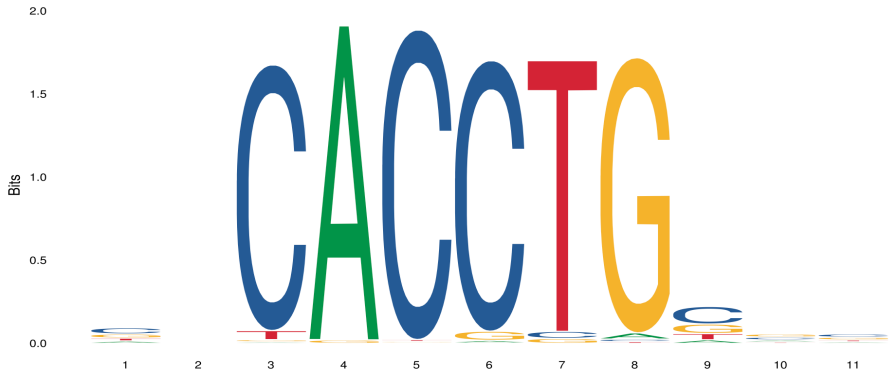
Walter+Eliza Hall
Institute of Medical Research
DISCOVERIES FOR HUMANITY

Surviving through Milton evolution:

From motif scanning to parallel processing of NGS data



Determining regulatory networks through motif scanning



A [	3211	4685	114	19905	32	193	4	441	1982	4258	3490	**]**
C [	7998	5550	19195	0	19878	19257	501	178	9561	6117	6955	**]**
G [	5228	5179	158	223	97	649	355	19381	5178	6601	5828	**]**
T [	3692	4715	662	1	122	30	19269	129	3408	3153	3856	**]**

TF motif finding

- **Tools:** MEME, Homer, clover
  - Use classical statistical tests, such as **log-likelihood ratio** or **hypergeometric/binomial** test , to find over-represented sequences given a Position Weight (PWM) or Position Frequency Matrix (PFM)
  - Recent approaches include **Neural Networks** , **Deep Learning** and **Graphical Models**
- **Motif databases:** JASPAR, HOCOMOCO, in-built databases and **TRANSFAC**
- **Processing time:**
  - **3 DAYS** for genome-wide motif scanning for a single TF on a Unix machine. These tools are mostly inherently single-threated.
  - **1 DAY** (on average) on **Milton** using a 'submit' queue with 1 node, 2 threads and 128gb memory if genome-wide, 64 or less if scanning targeted regions such as promoters only. And because we can run the jobs in parallel, we would **have results for all TFs in at most a couple of days.**

- Acknowledgments: Even, Daniel Cameron & Miguel/Steph.

Then, we started to receive many NGS data…

- Mostly RNA-Seq but also ChIP-Seq and ATAC-Seq to a lesser extend
- Unlike the motif scanning jobs, these required as many CPUs as possible, and a relatively fixed memory.
- Used **Milton** to process FASTQ files in parallel. This involved typical tasks such as read trimming, alignment, marking duplicate reads, sorting, indexing, peak calling etc
- Jobs were mostly submitted to 'medium' queues
- Personal queues couldn't really work for me, as the number of FASTQ files I receive and the processing steps that are involved are different from one project to the other.

However,

- By this point of the time many more people started to use Milton, or there were many more long running jobs holding up the resources, and it could take more than half a day for my routine jobs to start. Fortunately, the Milton team introduced **'static'** queues, which were just appropriate for the purpose of routine tasks I needed to do.

Milton speeded up some of our rather urgent tasks/analysis

An email from a collaborator:

"He had 4 of his own samples in the Seq line that he has some urgency to analyse. He can get the analysis he needs by himself, but he need some help for aligning his data, which I understand that can be done overnight in the computer. I wonder if you could make this favor for him? "

Thanks Milton team

- Over 13 projects that I have been involved in this year couldn’t have been possible without your good work on Milton development.