BINF sights.

ISSUE15
2022-T1-W8

humanity's missing piece

the secrets behind 'junk' DNA that impact our lives --- and our futures

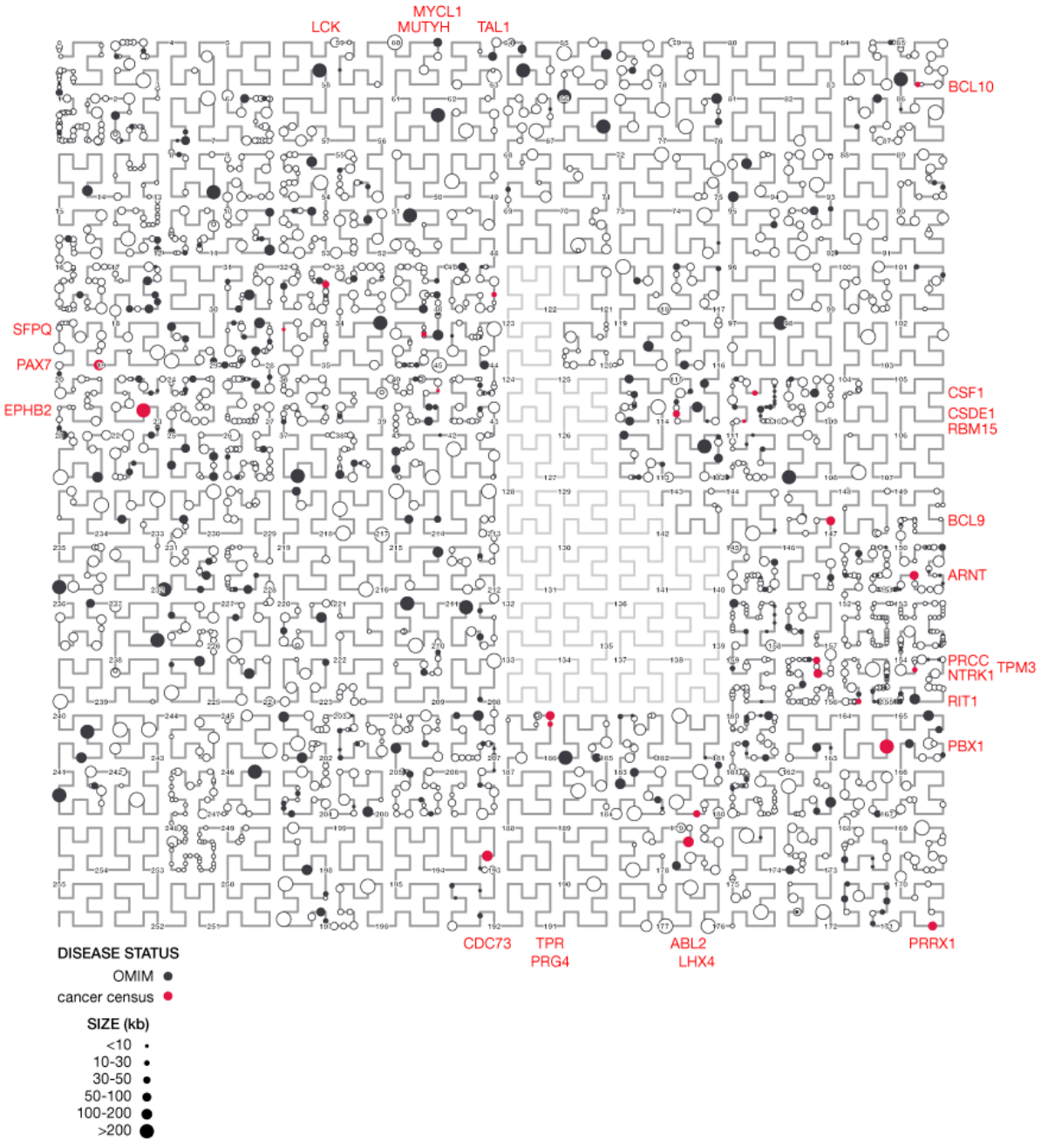


IMAGE SOURCE:
Martin
Krzywinski

Society News

BINFSOC

BINFsights Podcast.

In the coming weeks, the BINFSOC team will be launching a new platform for our bioinformatics content -- our very own podcast. The BINFsights Podcast will discuss various hot topics in the bioinformatics and biotechnology spaces, with a focus on more specific research areas from some of our leading academics at UNSW and beyond.

We'll also have guests to talk through advice for undergraduates, different career paths through research and industry, and balancing work and life outside BINF. We'll have more details coming soon on our social media, but in the meantime, if you have a burning question or a suggestion for a podcast topic (or if you'd like to be interviewed as a guest on the podcast!), don't hestitate to reach out to us at exec@unswbinfsoc.com

“

Eventually we’ll be able to sequence the human genome and replicate how nature did intelligence in a carbon-based system.

— Bill Gates

Opportunity:

Biomedical / Software Engineer.

nthalmic

Final year part-time / graduate roles

Biomedical/Software Engineer
A biomedical/software engineer with an aptitude to develop specialised software and a strong commitment to systematic documentation. The main purpose of the job is to develop technical software in the following areas: designing next-generation ophthalmic devices, supporting software infrastructure to facilitate advanced modelling of ophthalmic devices, and optimisation. Assist nthalmic's optical engineers by providing various software modules that perform functions within a broader framework of product developments.

more information: https://unswbinfsoc.com/nthalmic-opportunity

a: 2A Lord St, Botany NSW 2019
p: (02) 9037 7700

SEND RESUME / COVER LETTER

APPLY > hr@nthalmic.com

humanity's missing piece

WHAT WAS ONCE THOUGHT OF AS 'MISSING' OR 'JUNK' DNA MAY IN FACT BE A CRUCIAL PART TO OUR BLUEPRINT -- AND MAY EVEN HELP SOLVE THE PROBLEM OF AGING ITSELF.

Writer *Tom Parish* ***Editor*** *Cam McMenamie*

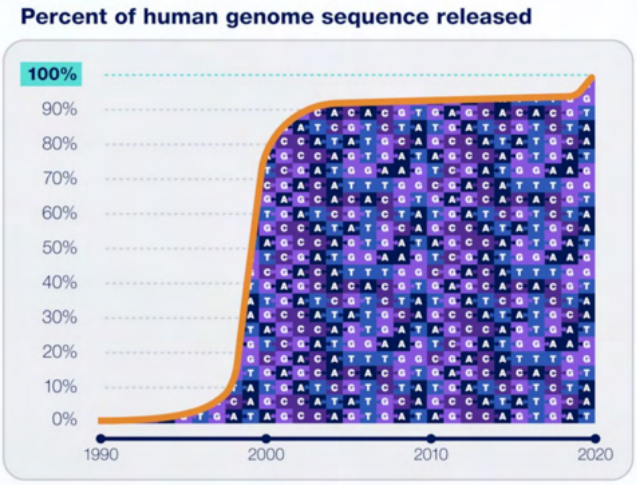
a piece of the jigsaw puzzle.

THE HUMAN GENOME PROJECT (HGP) was an internationally collaborative project with the goal of mapping out the entire human genome. The multi-billion dollar project was deemed complete on April 14, 2003, having sequenced 99% of the euchromatic genome, and around 92% of the entire genome [1]. This reference genome was revolutionary to the biological sciences, allowing us to more closely study important genes and their relationship to biological function and disease. The reference genome has also given scientists invaluable knowledge that led to the understanding of regions of our DNA that do not code for any proteins. These DNA sequences were often thought of as "junk", or DNA that no longer had a use (perhaps some function that existed in an ancestor that was removed via evolution).

Unexplored territory

However, in the 20 years since the project, these regions have been discovered to have important regulatory functions; can code for pseudogenes; make up telomeres and centromeres; as well as other important components. However, the final 8% has been puzzling scientists for years with very little progress being made in almost two decades (see picture). In fact, the most recent reference genome (GRCh38) has 151 mega–base pairs of unknown sequence distributed throughout the genome [2]. It wasn't until the Telomere-To-Telomere project started to venture into this unknown that we've been able to see a complete picture of these once uncategorised and unexplored regions.

But why has this taken so long? There are regions of the genome that have long been thought of as unsequencable, at least with the technological capabilities at the time. These regions are extremely lengthy stretches of DNA, sometimes hundreds of thousands of bases long [4], with short repeated subsequences. While these sequences might appear simple and consist of repeating motifs, the nature of their function has long been misunderstood as their exact make-up is unknown. These regions, called satellites, often found in the telomeres and centromeres of a chromosome, have important functions that are still being explored. Satellites were impossible to sequence at the time of the HGP due to the limitations in sequencing technology, with individual reads at the time only being around 500 bases long.



Graph from National Human Genome Research Institute showing the development of the human genome in the years since the HGP [3]

As you might expect, trying to determine the length of these repeats is impossible when you can only sequence a fraction of the total length of the region. The only real way to determine the nature of these areas of the genome was with significantly longer continuous reads. This is because shorter reads do not contain parts of the

sequence that overlap with the previous (or subsequent) sequences, and the number of subsequence repeats between two known locations is ambiguous.

New technology

So, what happened to make this possible? It wasn't until recently that sequencing technology had advanced to the degree that researchers could obtain reads that covered an entire repeated region. Oxford Nanopore sequencers have been able to do continuous reads over 2 million bases long in recent years [5], allowing DNA reads to span the entirety of these problematic regions. The information from the Nanopore sequencing, which unfortunately has high error rates of up to 5%, was combined with reads from PacBio sequencers which produce multi-kilobase reads with much lower error rates. These two technologies have allowed scientists to reassemble the sequencing data in a way that hasn't been possible before.

Findings

But why is this important? One may think that surely the regions that code for proteins are far more important than this satellite DNA, especially given that proteins are the functional units of any biological cell (think back to the 'central dogma' of introductory biology classes). Whilst it is true that we don't know as much about these regions as we do about other parts of the genome, this is because we have been missing the full picture. This new reference genome could explore the hypothesised impacts these regions have on an individual's predisposition to certain diseases. One such disease is Huntington's disease, a genetically inherited neurodegenerative condition that largely affects motor coordination and other functional brain abilities [1]. It might also allow insights into how telomeres (repeated DNA sequences at either end of a chromosome) work. These regions protect the more 'important' DNA during the replication process, where sometimes the ends of the chromosome are lost. Over time, after many generations, cells lose more and more of this information through replication. The telomeres thus serve as a 'buffer' for this effect, and have been compared to the ends of your shoelaces. This process is thought to be one of the many reasons for biological ageing, and future research might lead to new discoveries on slowing, or even reversing ageing. It also may provide insights into the function of satellites in centromeres, DNA regions that are thought to be crucial to the structural components that facilitate replication within a chromosome.

Overall these findings are just one piece of many that may help unlock the blueprints for future experiments, and may someday contribute to finding a 'standard model' of biological function in a similar way to what we have in physics.

If you'd like to read more about this research, you can find more information here [6]

References

[1]https://www.pbs.org/newshour/science/how-scientists-finally-completed-the-human-genomic-puzzle

[2]https://www.science.org/doi/10.1126/science.abj6987

[3]Picture:https://twitter.com/genome_gov/status/1415692505832583168/photo/1

[4]Satellite_length:https://genomebiology.biomedcentral.com/articles/10.1186/gb-2003-4-5-214

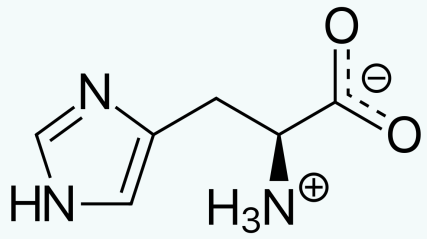
[5]https://nanoporetech.com/about-us/news/longer-and-longer-dna-sequence-more-two-million-bases-now-achieved-nanopore

[6]https://sites.google.com/ucsc.edu/t2tworkinggroup/home - site from the T2T consortium

AMINO ACID OF THE WEEK

[HISTIDINE]

CHEMICAL STRUCTURE



HISTIDINE

H

His

155.1546

DNA CODONS C A Y

POLARITY:
DEPENDS ON pH

POSITIVELY CHARGED AT PHYSIOLOGICAL pH

DISCOVERY:
ISOLATED IN 1896 BY PHYSICIAN Albrecht Kossel AND Sven Gustaf Hedin

PROTEINOGENIC - BUILDING BLOCK OF PROTEINS.

PRECURSOR TO HISTAMINE, IMPORTANT AS AN INFLAMMATORY AGENT IN THE IMMUNE SYSTEM.

CAN FORM COMPLEXES WITH MANY METALLIC IONS

IMIDAZOLE SIDECHAIN SERVES AS A LIGAND IN METALLOPROTEINS, NOTABLY ATTACHING TO Fe (IRON) IN HEMOGLOBIN (PROTEIN FOR OXYGEN TRANSPORT IN RED BLOOD CELLS)

BINF sights.

Contact us

-

IF YOU HAVE ANY COMMENTS or feedback regarding BINFsights, please write to us at binfsights@unswbinfsoc.com

We also encourage anyone to share with us anything you'd like us to take a look at, be it a bioinformatics tool that you have made or find useful; or news in the bioinformatics world that you'd like to see written about in future issues.

TO VIEW PAST AND PRESENT issues of BINFsights, check out our website at unswbinfsoc.com/binfsights

Stay tuned on our Facebook page for updates regarding events and society news.

-- The BINFSOC Team

BINF sights.