Euplotid

Author:
Diego Borges-Rivera

August 18, 2017

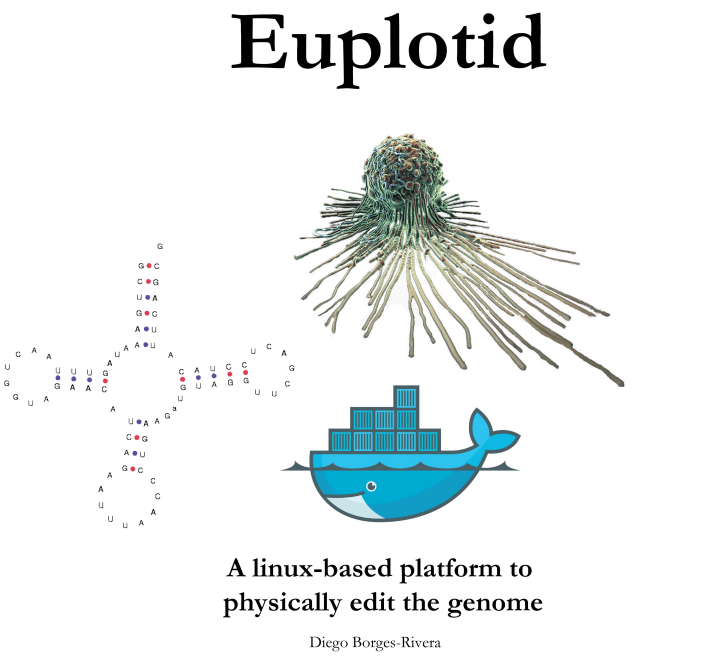

Figure 0.1: Graphical Abstract

1 Euplotid

1.1 Epigraph: The Naming of Euplotid

"The cell is a system that is capable of creating and conserving its own code" --Marcello Barbieri

The Cambrian explosion is responsible for the cementing of animal life on this planet; phylogenetic analysis appears to indicate that all metazoans originated from a single common flagellated organism, something resembling a Euplotid. Darwin has noted that this event does not seem to follow with a traditional evolutionary view, the speed and emergence of metazoans accelerated orders of magnitude when compared to the life that preceded them. There have been many attempts to explain this event, from Oxygen concentration to complexity thresholds, but no-one has considered that clues may lie within the genetic code itself, after all, this code is fundamental for the evolution of life as we know it.

A number of patterns have been observed within the codons, but a particularly interesting combination of two rules allows for the writing of the Euplotid genetic code in a contracted manner. By combining Rumer's Bisection (a)[1] with Hisagawa-Minata's[2] ordering by codon redundancy/mass (b) and applying them to the Euplotid genetic code we are able to contract the codons as first shown in Makukov et al[3].

Makukov et al discovered the protacted Euplotid codon arrangement and go well into the arithmetic interpretation, that is to say the part dealing with arithmatic. Below is an ideographical interpretation:

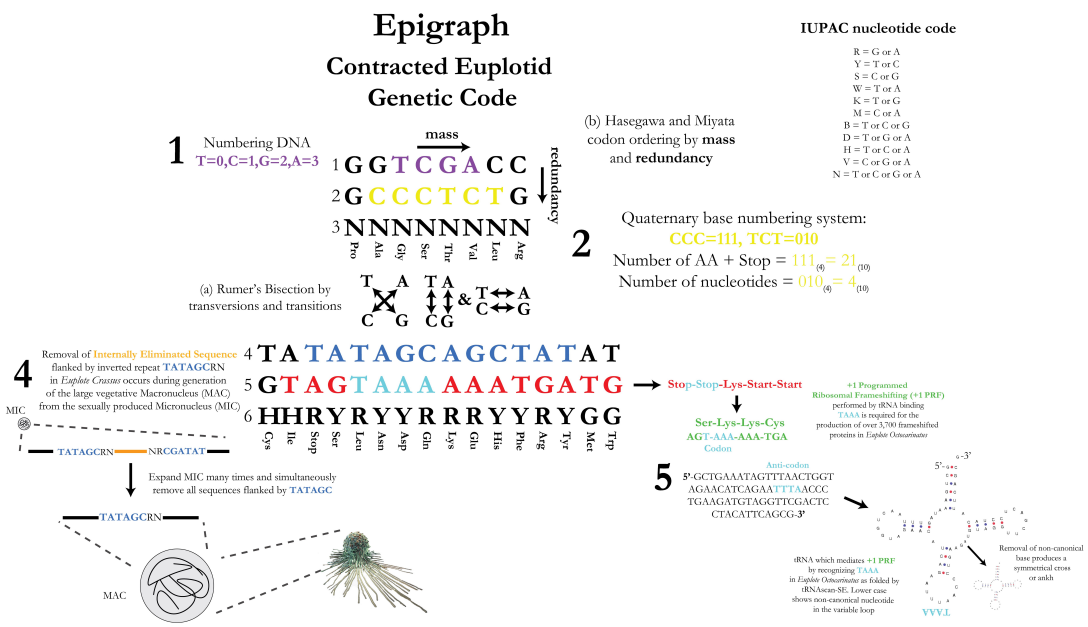

Figure 1.1: Contracted codons of the Euplotid Genetic code

1. GGTCGACC: How to number DNA, 0=T, C=1, G=2, A=3
2. GCCCTCTG: What radix number system to use. CCC = 111 (base 4) = 21 (base 10) = number of amino acids + stop codon. TCT = 010 (base 4) = 4 (base 10) = number of nucleotides.
3. NNNNNNNN: Is able to complement "GGTCGACC"

4. TATATAGCAGCTATAT: Euplotids remove Internal Eliminated Sequences (IESs) in the process of generating the large vegetative Macronucleus (MAC) from the small sexually produced Micronucleus (MIC). The Euplote Crassus consensus sequence is 5′-TATrGCRN-3′[4].
5. GTAGTAAAAAATGATG: Translating from left to right starting w/ TAG would produce: Stop-Stop-Lys-Start-Start using the canonical genetic code, but in Euplotids +1 Programmed Ribosomal

Frameshifting (+1 PRF) occurs at AAA sites preceding stop codons due to a tRNA recognizing UAAA instead of UAA[5] [6]. So if we perform a +1 PRF and read starting at AGT we get: Ser-Lys-Lys-Cys. This is only in Euplotids due to TGA coding for Cys instead of stop. +1 PRF in Euplotids is used to generate functional proteins from fusing different reading frames in over 3,700 proteins, including the Reverse Transcriptase of LINE-elements, ORF2[7]

6. HHRYRYYRRRYYRYGG: Is able to complement "TATATAGCAGCTATAT"

Due to this unique genetic signature, I decided to name this linux-based platform "Euplotid".

1.2 Abstract

1.2.1 http://dborgesr.github.io/Euplotid/

1.2.2 http://www.biorxiv.org/content/early/2017/08/14/170159

Euplotid is composed of a set of constantly evolving bioinformatic pipelines encapsulated and running in Docker containers enabling a user to build and annotate the local regulatory structure of every gene starting from raw sequencing reads of DNA-interactions, chromatin accessibility, and RNA-sequencing. Reads are quantified using the latest computational tools and the results are normalized, quality-checked, and stored. The local regulatory neighborhood of each gene is built using a Louvain based graph partitioning algorithm parameterized by the chromatin extrusion model and CTCF-CTCF interactions. Cis-Regulatory Elements are defined using chromatin accessibility peaks which are then mapped to Transcription Start Sites based on inclusion within the same neighborhood. Convolutional Neural Networks are combined with Long-Short Term Memory in order to provide a statistical model mimicking transcription factor binding, one neural network for each transcription factor in the genome is trained on all available Chip-Seq and SELEX data, learning what pattern of DNA oligonucleotides the factor binds. The neural networks are then merged and trained on chromatin accessibility data, building a rationally designed neural network architecture capable of predicting chromatin accessibility. Transcription factor binding and identity at each peak is annotated using this trained neural network architecture. By in-silico mutating and re-applying the neural network we are able to gauge the impact of a transition mutation on the binding of any human transcription factor. The annotated output can be visualized in a variety of 1D, 2D and 3D ways overlaid with existing bodies of knowledge, such as GWAS results. Once a particular CRE of interest has been identified by a biologist the difficulty of a Base Editor 2 (BE2) mediated transition mutation can be quantitatively assessed and induced in a model organism.

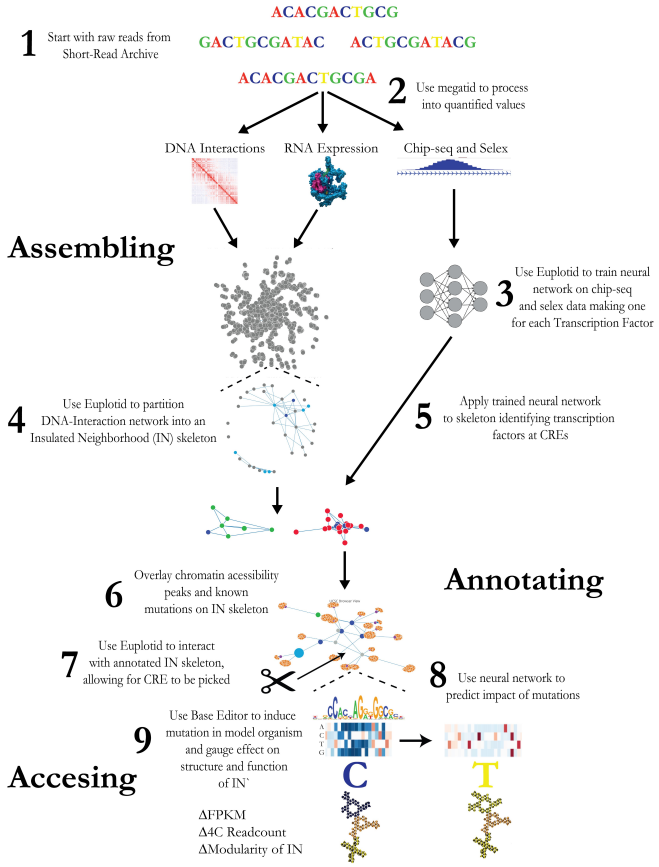

Figure 1.2: Detailed Abstract

1.3 Introduction

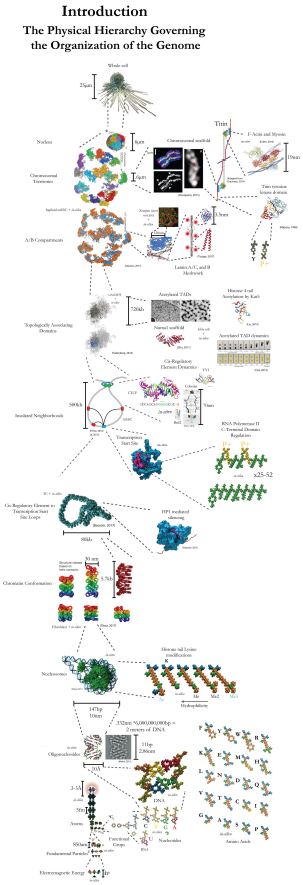

Figure 1.3: Introduction to the physical hiearachy organizing the genome. Deuterium Autocad model and brief movie describing it. 3D printable Autocad model Print using any 3D printer by printing 3 STL files, 1 Proton, 1 Electron, 1 Neutron, then using a paperclip as the axis of the electron attach it to the neutron. Use these craft magnets to attach the pieces together. Atomic Autocad model and brief movie describing it. Nucleotide Autocad model and brief movie describing it. DNA Autocad model and brief movie describing it. Amino Acids Autocad model and brief movie describing it. Nucleosome Autocad model and brief movie describing it. Histone Tail Autocad model and brief movie describing it. Hp1a mediated chromatin Autocad model and brief movie describing it. Chromatin Autocad model and brief movie describing it. C-Terminal RNA Polymerase Autocad model and brief movie describing it. Cis-Regulatory Element Autocad model and brief movie describing it. CCCTC-Binding Factor Autocad model and brief movie describing it. Cis-Regulatory Element – Transcription start site Autocad model and brief movie describing it.

The physical hierarchy of the genome begins with the most simple building block, what we previously though of as the indivisible unit we call elements. The history of elements harps back to 360 B.C. when the philosophy behind the physicality of elements was established[8], as exemplified in the Platonic Solids. The philosophical groundwork was laid for the understanding that compounds were made up of space-filling elements, and that through their combination one could create new compounds with very different properties than the sum of their parts. In the early 1800s the first atomic model was developed, able to explain chemical reactions as physical rearrangement of indivisible atoms[9]. The indivisibility of this atom was challenged in 1897 when the electron was discovered; the so called "plum pudding" model was born[10]. The plum pudding lasted until 1911 when the infamous gold

foil experiment proved that the positively charged "pudding" was actually a nucleus[11]. This nucleus contained protons, and later, was found to contain neutrons as well. Although the presence of the electron can be measured all around the positively charged nucleus they appear to be present at higher likelihoods in certain locations. During the 1930s a quantum electro dynamic model was born which is able to predict the probability of observing electrons at specific locations perfectly[12]. In tandem, particle physics gave us a clearer picture of the nucleus, as a tightly packed ball of protons and neutrons, each made up of quarks. This left us with the atom as a tightly packed nucleus surrounded by a field of electron "probability", doomed to never truly ever know exactly what will happen. In 2012 an interesting proposition was put forth, what if a quanta of energy itself had a physical shape, albeit 2D?[13] This gave rise to a novel way of interpreting all the previously developed theories, and coincidentally, loops back to Plato's first Platonic solid, the tetrahedron.

Around 1860 the philosophy was developed in order to explain the natural evolution which gave rise to the diversity in organisms as seen today[14]. The concept that all organisms come from a single common ancestor is in some ways disturbing in its simplicity, but is a key insight in order to understand our world as a whole. At the same time the physical explanation behind the tree of life was being laid down through the use P. Sativum[15]. The understanding that traits are inherited was extended to human disease in 1908 through the study of alkaptonuria[16]. Although the mechanism of inheritance was established, the physical material encoding the instructions for these traits was still hotly debated, was it proteins or nucleotides? The debate was settled in the 1930s through the use Pneumococcus and its virulence as a phenotypic trait[17]. With the chemical composition of the transforming material settled as nucleotides, the code defining the transition from DNA to RNA and then Amino Acids was solved during the 1950s[18]. Although we had found the chemical composition of the information storing component, we knew almost nothing as to how its shape was used to encode information. Two rules were discovered during the 1950s which began to decode DNA, Chargaff's rules. The first rule laid the groundwork for the structure of DNA to be solved, that is to say %C=%G and %A=%T[19]. DNA's structure was famously solved in 1953, giving a physical explanation to Chargaff's first rule and a major step in the physical organization of the genome was solved, the 3D shape of an oligonucleotide[20]. It is very interesting to note that although a physical explanation was found for Chargaff's first rule, his second remains unexplained.

The first amino acid was discovered in 1806, asparagine. Over the next decade the rest of the canonical 20 amino acids were discovered, isolated, and their properties carefully measured[21]. The amino acids form the functional unit of peptides, which when strung together and folded, create proteins. Our understanding of how these mechanical subunits come together is still in its infancy, in much due to our misunderstanding of the charge and mechanics at the most fundamental level. With a sharply defined boundary between energy and matter we are able to model interactions between these small protein building blocks in a far more natural, newtonian manner, while mantaining quantum accuracy.

Much of the confusion between the carrier of genetic information was due to DNA's extremely tight association with positively charged protein complexes called the nucleosomes[22]. The nucleosome's components and structures were developed and refined through the 1940s and 50s, with the core nucleosome's structure and components resolved at near atomic resolution. Although the predominant components were solved, new variants of the nucleosome complex were discovered, and continue to be, such as the newly characterized MacroH2A variants[23]. In the 1960s post-translational modifications of the nucleosome's tail were discovered to affect its association with DNA through changing lysine's charge and shape[24]. This level of the genome's physical organization allows for about 200bp to be neatly packaged, tagged and accessed, laying the groundwork for larger diameter fibers.

Chromatin can be understood as any shape of DNA that has a diameter larger than the canonical 10nm beads on a string nucleosome model. The exact shape and the in-vivo existence of the 30nm chromatin fiber has been hotly debated. It appears that there exists evidence for both sides, and in reality, it seems likely that the chromatin is a dynamic fiber, capable of deforming, memorizing, and reacting[25]. Within the last decade we have begun to probe how the shape and regulation of this fiber can impact its shape and function, namely we have just begun to unravel the consequences of histone tail marks on the conformation of chromatin. Seemingly, chromatin is much more than the sum of its parts, and a key way of maintaining information in the shape of our genome[26].

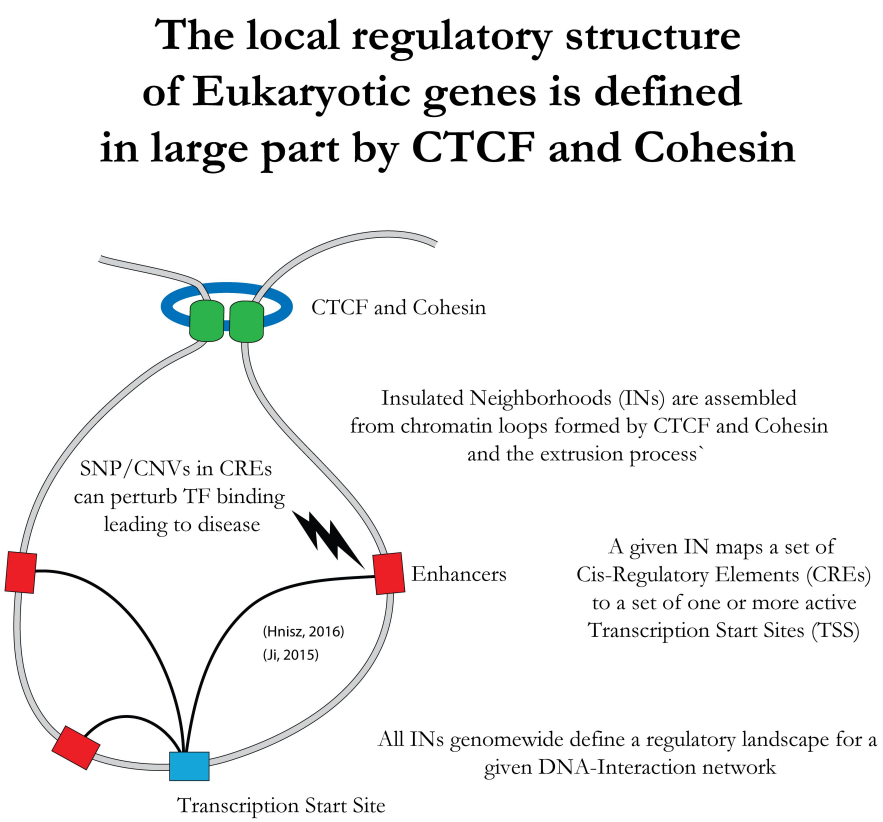

Figure 1.4: The local regulatory structure of Eukaryotic genes is largely maintained by clever use of CTCF and Cohesin whose misregulation can lead to disease

Once we had conclusively settled on DNA as the transforming material we could begin to elucidate the set of ordered reactions which takes DNA and decodes it to amino acids through an RNA intermediate. Throughout the 1970s the first steps of RNA-Polymerase mediated transcription were solved and the key players identified[27]. It was quickly discovered that not all RNA is destined for translation, only the subset coined "messenger RNA" or mRNA. RNA-Pol II was shown to be the holoenzyme responsible for the polymerization of this mRNA, which is in the reverse complement of the DNA template. As the constituents discovered for the pre-initiation complex (PIC) grew during the 1980s, it was found to form around the canonical TATA DNA motif, coined the "TATA-box"[28]. This TATA box has a specific 3D shape which causes a bend of the DNA at approximately 80 degrees when bound by proteins called Transcription Factors (TFs). The Mediator protein "arm" complex is attached to this TATA DNA bend. Mediator, aided by Nipbl, allows for the threading of this bent DNA through a small protenaicous band structure known as Cohesin, thereby creating a small loop[29]. The elongation of RNA-Pol II is preceded by TFIIH mediated phosphorylation of the C-Terminal Domain (CTD) of RNA-Pol II[30]. How the phosphorylation of the CTD impacts its interaction with the extremely electronegative surface of RNA-PolII has not been studied at the quantum level.

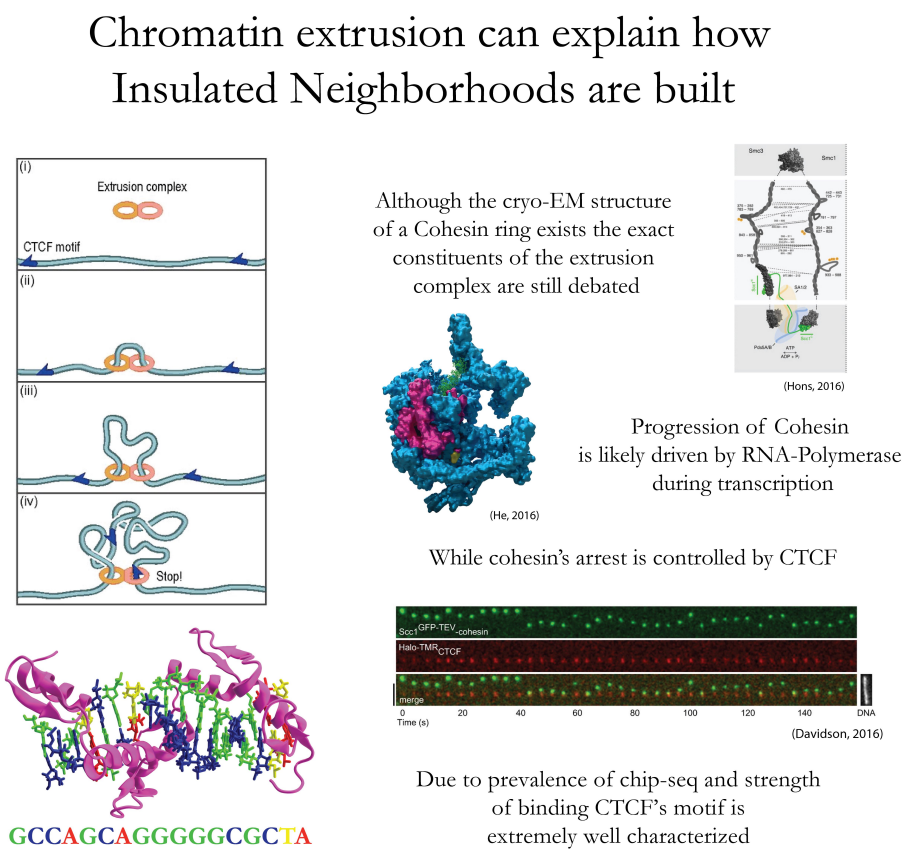

Figure 1.5: Chromatin extrusion allows for the stacking of Cohesin rings on CTCF

Recently it was found that the natural motor motion of RNA-Pol II during elongation serves to push this Cohesin ring, causing it to extrude a loop, bringing seemingly distant parts of the genome into direct physical proximity. This extrusion process continues until a bound CCCTC-Binding Factor (CTCF) is encountered, stopping the progression of Cohesin and causing the ring to stack at that bound CTCF site. CTCF's history in research took many turns, being assigned a plethora of roles, from enhancer to repressor, until its eventual establishment as a looping factor[31] . The rate of release of Cohesin at CTCF sites is also actively controlled by acetylation of the ring, while CTCF's binding can be impacted by the methylation of its DNA binding motif[32]. The intricate details are still debated, but overall it appears that the clever regulation of on/off rates on DNA for these three pieces, CTCF, Cohesin, and RNA-Pol II, allows for the creation of dynamic structures capable of reacting to differing cellular states, tuning a gene's local regulatory structure to adapt to specific environments. Although a CRE is able to influence the expression of any gene in extreme genomic proximity, the larger structures encompassing the CRE can cause it to impact TSSs from seemingly distant promoters[33]. The local regulatory structure of Eukaryotic genes is dependent on their own expression through the extrusion of Cohesin rings, this intriguingly forms a sort of feedback mechanism. The speed and dynamics of this transcriptional feedback mechanism may be influenced by certain charge dynamics from localized areas of Acetylation, such as those mediated by the H4 HAT complex or the asymmetrically loaded Acetylation of H3 at Super Enhancers[34]. Although these effects originate down from the very basic levels of the physical hierarchy, when aggregated together, it may be possible that they impact larger dynamics. We are beginning to see modeling approaches reaching the scales necessary to tackle chromatin looping questions, these models will continue to develop and gain in accuracy and generality. The physical hierarchy that controls the relation between CREs and their respective TSSs is a complex and extremely fine tuned process, but it may be that this complexity originates from very simple building blocks.

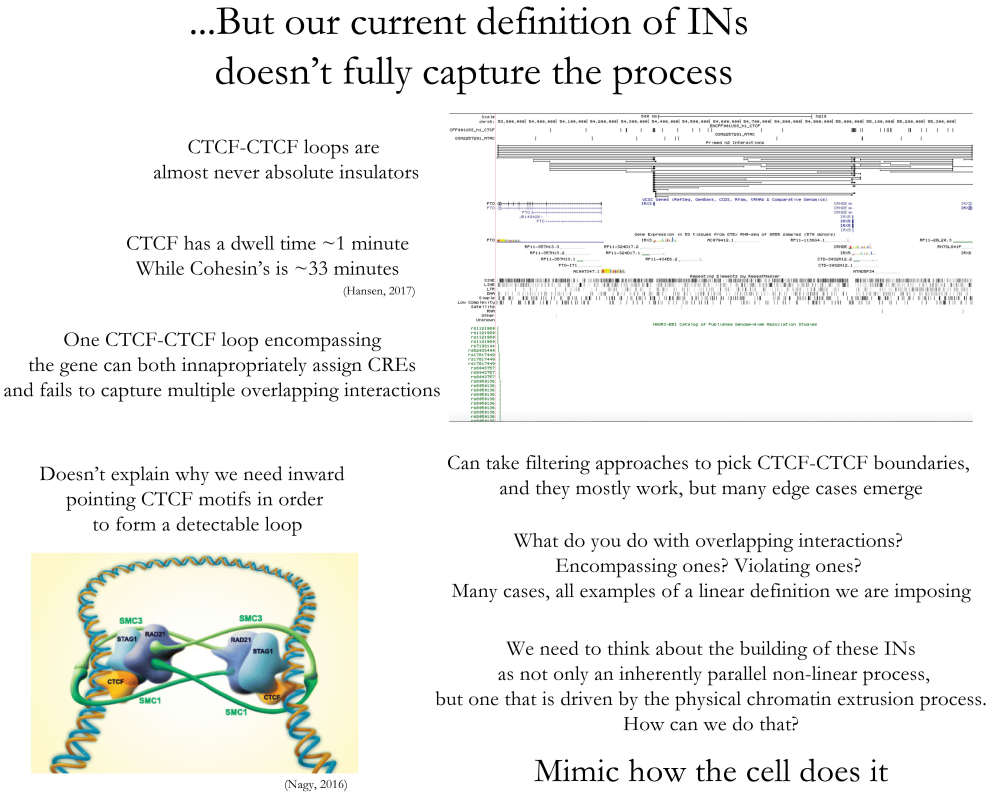

Figure 1.6: The definition of the local regulatory neighborhoods of genes is limited by linear thinking and genomewide measures

The regulation of transcription through the looping of CREs and TSSs appears to be often perturbed by disease causing mutations, especially those which are associated with non-coding CREs[35]. By combining recently developed methodologies to probe RNA-Seq, DNA-DNA interactions and Chromatin Accessibility with our recently acquired knowledge of the physical hierarchy governing the folding of the genome, we are able to build a rough picture of the 3D regulatory landscape of the genome. In order to digest this information in a manner which can guide experimentalists it is key to annotate and allow for easy access of these structures. Taking advantage of a number of recent developments in unrelated fields we are able to do just that; we provide a constantly evolving platform capable of allowing biologists to make physically informed decisions as what variation is causing the phenotype based on quantumly accurate models, virtually anywhere, anytime.

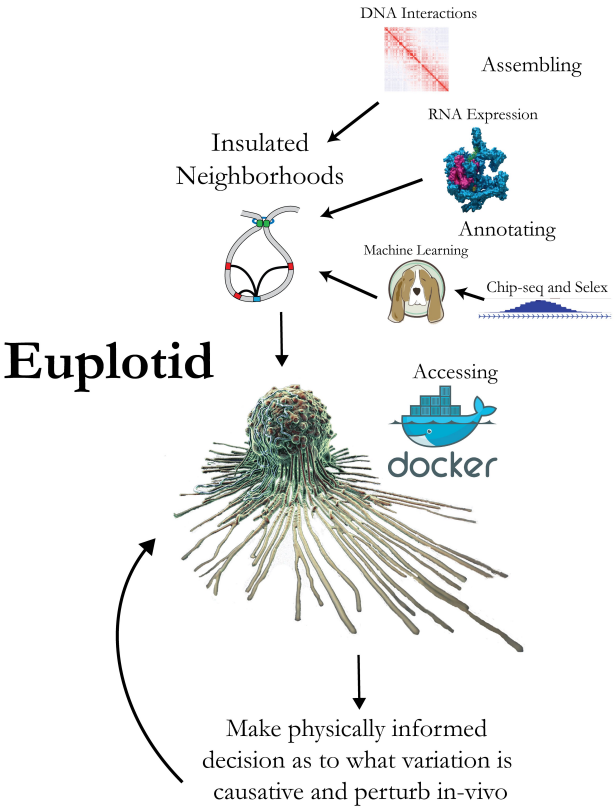

Figure 1.7: Euplotid solution

1.4 Results

1.4.1 Assembling

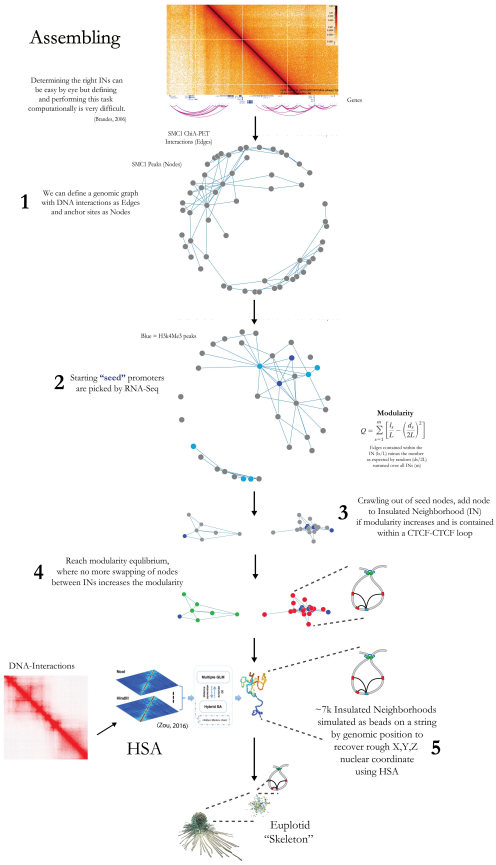

Figure 1.8: Assembling the local regulatory structure of genes

Determining the right local regulatory structure of a gene can be easy by eye but defining and performing this task computationally is very difficult.

Define Graphs We begin with set of Nodes, defined as a DNA range, with a left and right boundary, for example: chr16:55155024-53806737. We then add a set of Edges, defined as DNA-DNA interactions, or loops. These loops can be recovered from living cells using a variety of methods, such as Hi-C, In-situ Hi-C, ChiA-PET, HiChIP, GAM, etc, each having their own wet and dry processing protocol, with all dry implemented within Megatid.

Define starting nodes The initial starting conditions are when every Node is its own unique IN of size 1. We can use RNA-seq as a poor-man's proxy for the rate of Cohesin-RNAPolII mediated chromatin extrusion. The processing of raw RNA-Seq reads can be quickly performed within Megatid using STAR to align the reads and RSEM to quantify RPKM[36] [37]. We begin the algorithm by sorting the starting nodes by RPKM.

Crawl outward adding nodes Beginning at each highly transcribed Transcription Start Site we select the neighbor which would increase the overall network architecture the most as defined by Modularity and reassign its IN if the entire IN is encompassed within a single CTCF-CTCF interaction [38–40].

Reach modularity equlibrium Continue checking for valid reassignments until no more moves exist which increase the overall graph's modularity, thereby reaching an equilibrium.

Recover rough X,Y,Z location of IN Take all DNA-DNA interactions and combined with Lamin A/B1 Chip-Seq estimate the rough nuclear X,Y,Z position of each IN node by feeding the data through HSA [41]. The size of the node is defined using the sum of all reads falling within chromatin accessible regions.

1.4.2 Annotating

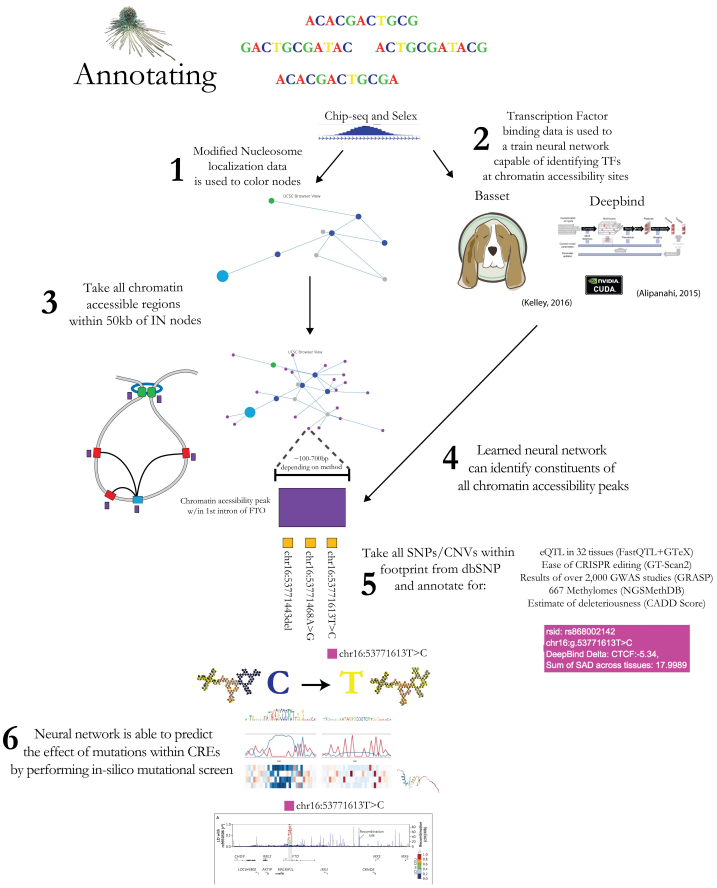

Figure 1.9: Annotating the Insulated Neighborhoods

Color nodes After assembling the Insulated Neighborhood skeleton it is key to be able to visualize the impact of Histone modifications on the local regulatory structure, therefore we simply color the nodes of the genomic graph by histone modifications in the given cell state of interest. Specifically, Red if node overlaps only with H3K27Ac and blue if it overlaps H3k4me3 [42].

Train neural networks Begin by training Convolutional Neural Networks (CNNs) based on all chip-seq and SELEX data for all TFs ever surveyed. The initial implementation of Euplotid uses pre-trained CNNs from Deepbind [43]. These CNNs are able to identify the TFs which fall under each chromatin accessiblity peak, but in order to understand the peak as a whole Euplotid takes advantage of Basset to train neural networks which are capable of predicting changes in chromatin accessibility [44]. Basset is trained on all available chromatin accessibility data in ENCODE, DNAse of 180 different cell lines. Basset is therefore able to perform in-silico simulations to gauge the impact of a given mutation on the the complex as a whole (SNP Accessibility Difference (SAD) profile), by combining this with the CNNs from Deepbind, we are able to make a prediction as to what factor is causing this change.

Select chromatin accessibility peaks Taking all chromatin accessibility peaks within a set distance (50kb) from all the nodes of a given IN allows us to identify the relevant areas of chromatin which are actively being used in this particular cell state, some potentially acting as Cis-Regulatory Elements. Any method of chromatin accessibility is appropriate, DNAse-seq, ATAC-seq, MNAse-seq, etc, all can be used as inputs.

Identify TF constituents Applying the trained neural network on each chromatin accessibility peak we are able to identify the constiuents, thereby identifying complexes putatively making up each Cis-Regulatory element. Currently this is performed by Deepbind, but a custom built pytorch based network will soon be implemented [45].

Select and annotate SNPs/CNVs We can then include all SNPs/CNVs within dbSNP which overlap the chromatin accessibility peaks within the IN [46]. This variation is then annotated with the following data if available: eQTL in 32 tissues (FastQTL+GTeX), ease of CRISPR editing (GT-Scan2), results of over 2,000 GWAS studies (GRASP), 667 Methylomes (NGSMethDB), estimate of deleteriousness (CADD Score)

Predict effect of SNPs/CNVs For each variant which falls within the IN we perform an in-silico mutational analysis. This in-silico mutational analysis is simple, predict the chromatin accessibility with and without the variant. The difference in SNP accessibility (SAD score) is calculated by Basset with the pre-trained networks as described above.

1.4.3 Accessing

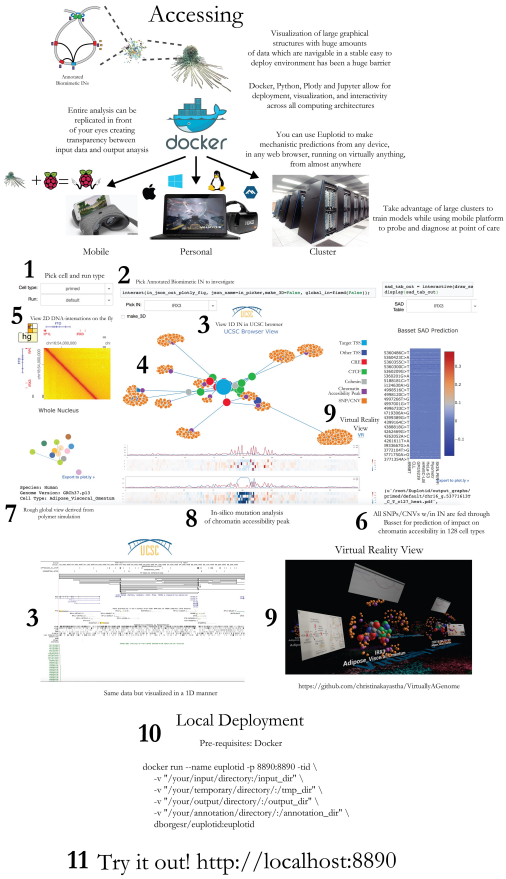

Figure 1.10: Accessing the Insulated Neighborhoods

Visualization of large graphical structures with huge amounts of data which are navigable in a stable easy to deploy environment has been a huge barrier. Docker, Python, Resin, Plotly and Jupyter together allow for deployment, visualization, and interactivity across all computing architectures [47] [48] [49] [50]. The entire analysis can be replicated and edited in front of your eyes, creating unprecedented transparency between input data and output analysis. You can use Euplotid to make mechanistic predictions from any device, in any web browser, running on virtually anything, from almost anywhere.

Pick cell type and condition Using widgets in Jupyter we are able to dynamically access the annotated local regulatory structure of every gene stored as graphs within JSONs through the backend. A Jupyter widget is a simple lightweight node.js wrapper to traditional python methods. Here we can pick what cell type and condition we want to investigate.

Pick annotated Insulated Neighborhood After picking a cell type and condition the list of annotated INs will be populated. A simple dropdown sorted by name is provided.

UCSC genome browser view If the user wants visualize the data in the traditional 1D manner it is possible to load the data into the UCSC genome browser. In this case we set the linear left and right boundaries as the leftmost and rightmost node within the IN currently being viewed.

Annotated Insulated Neighborhood The annotated IN is positioned according to the Fruchterman-Reingold force-directed algorithm on DNA-interaction read count in order to have a more visually pleasing view. By employing 3Djs and Plotly we are able to navigate large graphical structures with relative ease. When hovering over every DNA-Interaction node the following pieces of data are shown if available: eQTL in 32 tissues (FastQTL+GTeX), ease of CRISPR editing (GT-Scan2), results of over 2,000 GWAS studies (GRASP), 667 Methylomes (NGSMethDB), estimate of deleteriousness (CADD Score), and CNN prediction of TF identity

DNA-DNA interaction heatmap view Higlass.io Awesome tile-based viewing tool for DNA-DNA interaction data[51]. Employing D3.js to query a robust backend, this dockerized application is able to serve huge compressed multi-resolution 2D Hi-C data essentially instantly. The compression and raw interaction handling is done by Cooler.

SNP accessibility difference prediction Employing CNNs previously trained on Chip-Seq and SELEX data and combining them with LSTM networks we are able to predict the chromatin accessibility of a particular sequence in a given cell type. Taking all SNPs/CNVs which fall within the IN we then predict the impact of each of those on the accessibility of chromatin.

Global view HSA Using the X,Y,Z coordinates previously generated for each IN using HSA we are able to have a small "mini-map" corresponding to a global view of the entire nucleus. Each IN node is colored according to chromsome and the INs are connected according to genomic coordinate.

In-silico mutational analysis Basset Using previously trained neural networks we are able to view the predicted image for a given in-silico mutatation. This gives a quick and easy to assess view of the impact a given SNP has on a Cis-Regulatory element, potentially affecting its function.

Virtual reality view Taking advantage of Virtual Reality (VR) technology developed for both the military and consumer markets we are able to render the annotated INs in full immersive VR. We use Unreal Engine to design, build, and deploy the VR view[52]. Tested with the HTC Vive allowing for fully immersive room-scale exploration of large complex annotated INs. A more detailed explanation is available below

Deployment of Euplotid **INSTALL DOCKER HERE** Then open your terminal or cmd and: ~ docker run --name euplotid -p 8890:8890 -tid
-v "/your/input/directory:/input_dir"
-v "/your/temporary/directory/:/tmp_dir"
-v "/your/output/directory/:/output_dir"
-v "/your/annotation/directory/:/annotation_dir"
dborgesr/euplotid:euplotid ~

Try it out! `http://localhost:8890`

1.5 Discussion

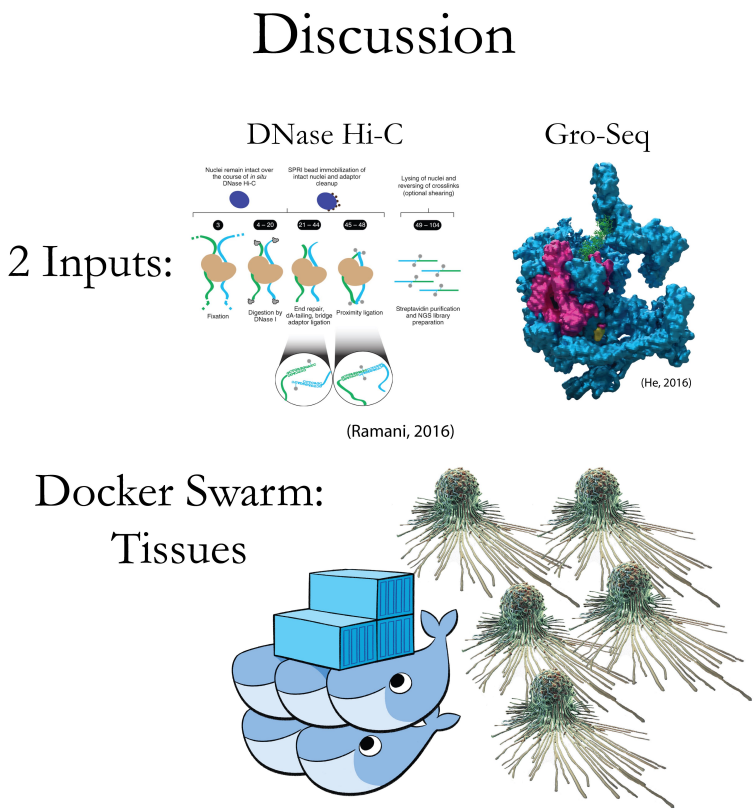

Figure 1.11: Next steps and impact

Euplotid is a linux based platform that is built to evolve over time, to shed pieces and gain new ones as more powerful bioinformatic tools are created. Due to its modularity and deployability Euplotid can be used almost anywhere. Combined with emerging "Edge" computing and sequencing infrastructures such as NVIDIA's Jetson and Oxford Nanopore's MinION[53] the potential for on-site building and visualization, from raw sequencing data to annotated immersive VR would be possible. A strategy which may be limited in immediate computing power but is able to deal with privacy concerns from the ground up.

In the future it will be possible to combine multiple images of Euplotid running in tandem mimicking tissues, with each Euplotid image communicating between each other and being slightly different, incorporating single-celled resolution techniques. Due to Euplotid's foundational principles we are able to capture movement and mechanics down to the quantum level but remain extremely efficient and tractable, we render what we need to look at. The availability and ease of use will allow Euplotid to spread around the globe with relative ease.

1.6 Acknowledgements

Acknowledgements

Rick
Eric
Dan
Charlie
Michael
Ben
Toni
Rest of the Young Lab

Figure 1.12: Acknowledgements

1.7 Methods

The process of building Euplotid began with taking raw sequencing reads stored in a few different formats and processing it all the way to quantified values. Due to the pliability, breadth, and flexibility of the methods they will be documented within their own Jupyter notebook. Acting as both the documentation and the pipeline itself this format allows for seamless data integration.

- Hello world intro to programming and Jupyter's capabilities helloWorld O*.
- Databases and good tools to crawl the internet for interesting datasets and hypothesis databasesTools O*.
- Fetch any type of sequencing data from SRA getFastqReads O
- QC, trim, and filter sequencing reads fq2preppedReads O
- Call peaks from Chip-Seq and Chromatin Accessibility reads fq2peaks O
- Call normalized interactions from ChiA-PET reads fq2ChIAInts O
- Call normalized Interactions from HiC reads fq2HiCInts O
- Call normalized interactions from Hi-ChIP reads fq2HiChIPInts O
- Call normalized interactions from DNAse-HiC reads fq2DNAseHiCInts O
- Call normalized expression and counts from RNA-Seq reads fq2countsFPKM O
- Call differentially expressed genes from RNA-seq counts countsFPKM2DiffExp O
- Call normalized counts, miRNA promoters, and nascent transcripts from Gro-Seq reads fq2GroRPKM O
- Call normalized interactions from 4C fq24CInts O

- Build, annotate and add INs to global graph for a given cell state using DNA-DNA interactions, Chromatin Accesibility, and FPKM addINs *
- View current built and annotated INs for all cell types viewINs O*.
- Search for and/or manipulate annotation and other data available to euplotid annotationManagement O*.
- Description of default software packages and images installed, how to get new ones, and which ones are currently installed. packageManagement O*.
- Find clusters of interconnected nodes (Communities) using a Louvain algorithm then visualize the results vanillaCommunities
- Create, manipulate, and visualize cool DNA-DNA interaction files chilledInteractions
- Design Base Editor and sgRNA plasmids for transition mutation at picked Cis-Regulatory Element CRE2plasmid

[O] = Megatid compatible [*] = Euplotid compatible [.] = Minitid compatible

1.7.1 helloWorld

Hello world intro to programming, ipython, and Euplotid

1.7.2 databasesTools

Databases and good tools to crawl the internet for interesting datasets and hypothesis. Some examples include GTeX, uniprot, SRA, GEO, etc, check them out!!

1.7.3 getFastqReads

Allows you to use Tony to find local fastq.gz files OR provide an SRA number to pull from

1.7.4 fq2preppedReads

Take fq.gz reads and QC them using FastQC checking for over-represented sequences potentially indicating adapter contamination. Then use cutadapt and sickle to filter and remove adapters. Can also use trimmomatic for flexible trimming.

1.7.5 fq2peaks

Take fq.gz align it using bowtie2 to the genome. Then using Homer software pick the type of peak (histone, chip-seq, dnase, etc) and chug through to get bed files of peaks. Can also use MACS2 w/ specific analysis parameters to deal with different types of peak finding problems.

1.7.6 fq2ChIAInts

Take fq.gz reads, prep them by removing bridge adapters (can deal with either bridges), align, find interactions, normalize, and spit into cooler format for later viewing. Can perform analysis using either Origami or ChiA-PET2

1.7.7 fq2HiCInts

Take fq.gz reads and chug them through HiCPro w/ tuned relevant parameters. In the end spits out a cooler file which can be loaded for further visualization.

1.7.8 fq2HiChIPInts

Take fq.gz reads and chug them through customized Origami pipeline and customized HiCPro pipeline. In the end spits out a cooler file which can be loaded for further visualization.

1.7.9 fq2DNAseHiCInts

Take fq.gz reads and chug them through HiCPro pipeline. In the end spits out a cooler file which can be loaded for further visualization.

1.7.10 fq2countsFPKM

Take fq.gz reads and chug them through STAR aligner and then RSEM pipeline. In the end spits out a counts vs transcripts matrix and a normalized transcript/gene FPKM matrix.

1.7.11 countsFPKM2DiffExp

Take RNA-seq count and FPKM matrix and run any one of many R packages (DESeq2,DESeq,EBSeq,edgeR...) to call differentially expressed genes. Plotting and interactive visualization of results included

1.7.12 fq2GroRPKM

Take fq.gz reads and align them using bowtie2 then find nascent transcripts using FStitch and miRNA promoters using mirSTP

1.7.13 fq24CInts

Take fq.gz reads and align them using bowtie2. Chug them through HiCPro and/or custom pipeline to get cooler file

1.7.14 addINs

Build, annotate and add INs to global graph for a given cell state using DNA-DNA interactions, Chromatin Accesibility, and FPKM.

1.7.15 viewINs

View current built and annotated INs for all cell types

1.7.16 annotationManagement

Search for and/or manipulate annotation and other data available to euplotid

1.7.17 packageManagement

Description of default image and the software packages that are installed, also how to get new packages, and how to export environment in yaml file for others to replicate analysis.

1.7.18 vanillaCommunities

Find clusters of interconnected nodes (Communities) using a Louvain algorithm then visualize the results

1.7.19 chilledInteractions

Create, manipulate, and visualize cool DNA-DNA interaction files

1.7.20 CRE2plasmid

Design Base Editor and sgRNA plasmids for transition mutation at picked Cis-Regulatory Element

1.8 Appendix: Biosemiotics in the Physical realm

1.8.1 Using DNA as a checksum

The ability to quickly verify accurate information transfer between a 1D DNA sequence and a 3D Matter topology is possible given certain assumptions hold: * Standard genetic code * Cytoplasmic pH conditions

1.8.2 Checksum functions

- DNA side (base 4)[3]:

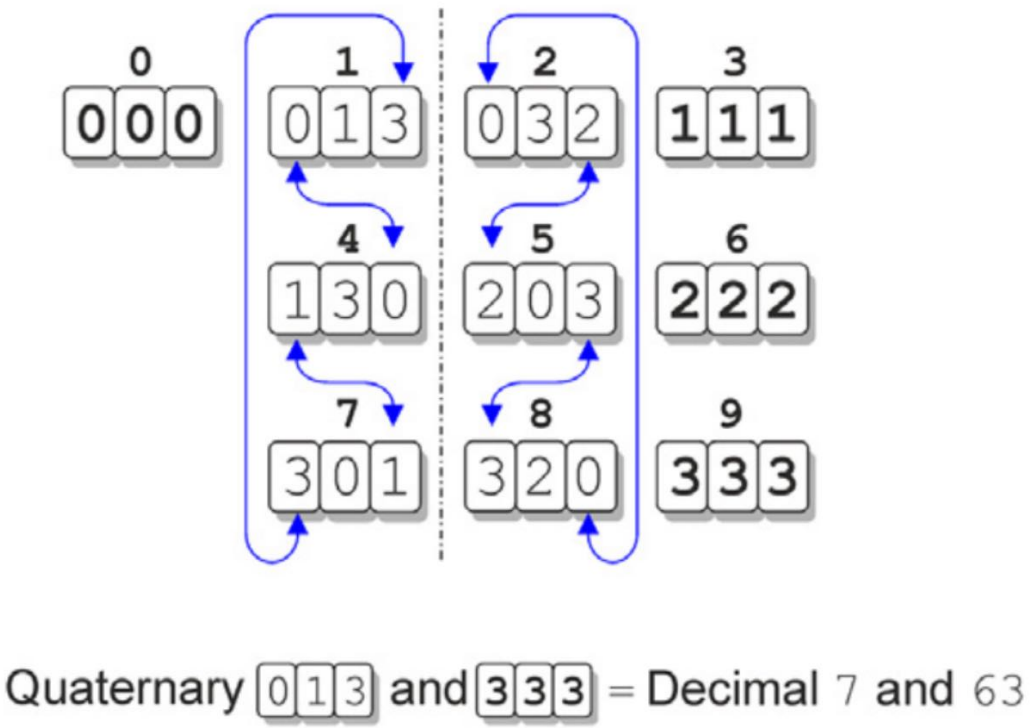

Figure 1.13: DNA numbering system, original figure from Makukov et al

- Amino Acid side (base 10)[3]:

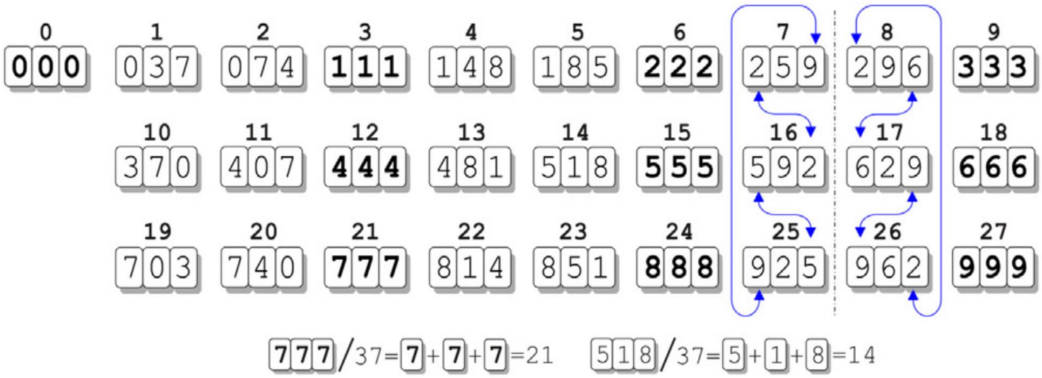

Figure 1.14: Amino acid numbering system, original figure from Makukov et al

The ability becomes useful as an indexing method to find errors when certain parameters are mapped to physical realities, namely the DNA backbone and the amino acid backbone.
* Z-DNA is 12=34 *bp per full turn*
Amino acids have 74=37*2 Deuteriums per backbone chain

You can start using the process of doing the checksum as a way to check for errors in physical reality. It remains an open question if a similar approach can be taken towards more complex systems, such as the cytoskeleton and signal transduction cascades.

2 References

[1] H.-J. Danckwerts and D. Neubert. Symmetries of genetic code-doublets. *Journal of Molecular Evolution*, 5(4):327–332, December 1975. ISSN 0022-2844, 1432-1432. doi:10.1007/BF01732219.

[2] Masami Hasegawa and Takashi Miyata. On the antisymmetry of the amino acid code table. *Origins of life*, 10(3):265–270, September 1980. ISSN 0302-1688, 1573-0875. doi:10.1007/BF00928404.

[3] Vladimir I. shCherbak and Maxim A. Makukov. The "Wow! signal" of the terrestrial genetic code. *Icarus*, 224(1):228–242, May 2013. ISSN 0019-1035. doi:10.1016/j.icarus.2013.02.017.

[4] John R. Bracht, Wenwen Fang, Aaron David Goldman, Egor Dolzhenko, Elizabeth M. Stein, and Laura F. Landweber. Genomes on the Edge: Programmed Genome Instability in Ciliates. *Cell*, 152 (3):406–416, January 2013. ISSN 0092-8674. doi:10.1016/j.cell.2013.01.005.

[5] Alexei V. Lobanov, Stephen M. Heaphy, Anton A. Turanov, Maxim V. Gerashchenko, Sandra Pucciarelli, Raghul R. Devaraj, Fang Xie, Vladislav A. Petyuk, Richard D. Smith, Lawrence A. Klobutcher, John F. Atkins, Cristina Miceli, Dolph L. Hatfield, Pavel V. Baranov, and Vadim N. Gladyshev. Position dependent termination and widespread obligatory frameshifting in Euplotes translation. *Nature structural & molecular biology*, 24(1):61–68, January 2017. ISSN 1545-9993. doi:10.1038/nsmb.3330.

[6] Ruanlin Wang, Jie Xiong, Wei Wang, Wei Miao, and Aihua Liang. High frequency of +1 programmed ribosomal frameshifting in Euplotes octocarinatus. *Scientific Reports*, 6, February 2016. ISSN 2045-2322. doi:10.1038/srep21139.

[7] Thomas G. Doak, David J. Witherspoon, Carolyn L. Jahn, and Glenn Herrick. Selection on the Genes of Euplotes crassus Tec1 and Tec2 Transposons: Evolutionary Appearance of a Programmed Frameshift in a Tec2 Gene Encoding a Tyrosine Family Site-Specific Recombinase. *Eukaryotic Cell*, 2 (1):95–102, January 2003. ISSN 1535-9778, 1535-9786. doi:10.1128/EC.2.1.95-102.2003.

[8] Donald Zeyl. Plato's *Timaeus*. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2014 edition, 2014.

[9] John Dalton. On the absorption of gases by water and other liquids. *Philosophical Magazine*, 24(93): 15–24, February 1806. ISSN 1941-5796. doi:10.1080/14786440608563325.

[10] Thomson J. Cathode Rays. *Philosophical Magazine*, 44(269):293–316, October 1897. ISSN 1941-5982. doi:10.1080/14786449708621070.

[11] E. Rutherford F. The scattering of α and β particles by matter and the structure of the atom. *Philosophical Magazine*, 92(4):379–398, February 2012. ISSN 1478-6435. doi:10.1080/14786435.2011.617037.

[12] E. Schrödinger. Quantisierung als Eigenwertproblem. *Annalen der Physik*, 384(4):361–376, January 1926. ISSN 1521-3889. doi:10.1002/andp.19263840404.

[13] Kelvin C. Abraham. An Introduction to Tetryonic Theory. *2014*, May 2014.

[14] Charles Darwin and Alfred Wallace. On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Journal of the Proceedings of the Linnean Society of London. Zoology*, 3(9):45–62, August 1858. ISSN 1945-9416. doi:10.1111/j.1096-3642.1858.tb02500.x.

[15] Scott Abbott and Daniel J. Fairbanks. Experiments on Plant Hybrids by Gregor Mendel. *Genetics*, 204(2):407–422, October 2016. ISSN 0016-6731, 1943-2631. doi:10.1534/genetics.116.195198.

[16] Garrod Sir Archibald. Inborn Errors of Metabolism. *American Journal of Human Genetics*, 10(1):3–32, March 1958. ISSN 0002-9297.

[17] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. *The Journal of Experimental Medicine*, 79(2):137–158, February 1944. ISSN 0022-1007.

[18] Arthur Kornberg, S. R. Kornberg, and Ernest S. Simms. Metaphosphate synthesis by an enzyme from Escherichia coli. *Biochimica et Biophysica Acta*, 20:215–227, January 1956. ISSN 0006-3002. doi:10.1016/0006-3002(56)90280-3.

[19] Erwin Chargaff. Preface to a Grammar of Biology. *Science*, 172(3984):637–642, May 1971. ISSN 0036-8075, 1095-9203. doi:10.1126/science.172.3984.637.

[20] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. J.D. Watson and F.H.C. Crick. Published in Nature, number 4356 April 25, 1953. *Nature*, 248(5451):765, April 1974. ISSN 0028-0836.

[21] E. Baumann. Ueber Cystin und Cystein. *Zeitschrift für Physiologische Chemie*, 8(4):299–305, 1883.

[22] Albrecht Kossel and others. Protamines and histones. 1928.

[23] Avnish Kapoor, Matthew S. Goldberg, Lara K. Cumberland, Kajan Ratnakumar, Miguel F. Segura, Patrick O. Emanuel, Silvia Menendez, Chiara Vardabasso, Gary LeRoy, Claudia I. Vidal, David Polsky, Iman Osman, Benjamin A. Garcia, Eva Hernando, and Emily Bernstein. The histone variant macroH2A suppresses melanoma progression through regulation of CDK8. *Nature*, 468(7327): 1105–1109, December 2010. ISSN 0028-0836. doi:10.1038/nature09590.

[24] V. G. Allfrey, R. Faulkner, and A. E. Mirsky. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS*. *Proceedings of the National Academy of Sciences of the United States of America*, 51(5):786–794, May 1964. ISSN 0027-8424.

[25] Sergei A. Grigoryev and Christopher L. Woodcock. Chromatin organization — The 30nm fiber. *Experimental Cell Research*, 318(12):1448–1455, July 2012. ISSN 0014-4827. doi:10.1016/j.yexcr.2012.02.014.

[26] Viviana I. Risca, Sarah K. Denny, Aaron F. Straight, and William J. Greenleaf. Variable chromatin structure revealed by in situ spatially correlated DNA cleavage mapping. *Nature*, 541(7636):237–241, January 2017. ISSN 0028-0836. doi:10.1038/nature20781.

[27] Maurice J. Bessman, I. R. Lehman, Ernest S. Simms, and Arthur Kornberg. Enzymatic Synthesis of Deoxyribonucleic Acid II. GENERAL PROPERTIES OF THE REACTION. *Journal of Biological Chemistry*, 233(1):171–177, January 1958. ISSN 0021-9258, 1083-351X.

[28] R. P. Lifton, M. L. Goldberg, R. W. Karp, and D. S. Hogness. The Organization of the Histone Genes in Drosophila melanogaster: Functional and Evolutionary Implications. *Cold Spring Harbor Symposia on Quantitative Biology*, 42:1047–1051, January 1978. ISSN 0091-7451, 1943-4456. doi:10.1101/SQB.1978.042.01.105.

[29] Michael T. Hons, Pim J. Huis in 't Veld, Jan Kaesler, Pascaline Rombaut, Alexander Schleiffer, Franz Herzog, Holger Stark, and Jan-Michael Peters. Topology and structure of an engineered human cohesin complex bound to Pds5B. *Nature Communications*, 7:ncomms12523, August 2016. ISSN 2041-1723. doi:10.1038/ncomms12523.

[30] Philip J. Robinson, Michael J. Trnka, David A. Bushnell, Ralph E. Davis, Pierre-Jean Mattei, Alma L. Burlingame, and Roger D. Kornberg. Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. *Cell*, 166(6):1411–1422.e16, September 2016. ISSN 0092-8674, 1097-4172. doi:10.1016/j.cell.2016.08.050.

[31] Jennifer E. Phillips and Victor G. Corces. CTCF: Master Weaver of the Genome. *Cell*, 137(7): 1194–1211, June 2009. ISSN 0092-8674, 1097-4172. doi:10.1016/j.cell.2009.06.001.

[32] Hideharu Hashimoto, Dongxue Wang, John R. Horton, Xing Zhang, Victor G. Corces, and Xiaodong Cheng. Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Molecular Cell*, 66(5):711–720.e3, June 2017. ISSN 1097-2765. doi:10.1016/j.molcel.2017.05.004.

[33] Denes Hnisz, Abraham S. Weintraub, Daniel S. Day, Anne-Laure Valton, Rasmus O. Bak, Charles H. Li, Johanna Goldmann, Bryan R. Lajoie, Zi Peng Fan, Alla A. Sigova, Jessica Reddy, Diego Borges-Rivera, Tong Ihn Lee, Rudolf Jaenisch, Matthew H. Porteus, Job Dekker, and Richard A. Young. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280): 1454–1458, March 2016. ISSN 0036-8075, 1095-9203. doi:10.1126/science.aad9024.

[34] Warren A. Whyte, David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell*, 153(2):307–319, April 2013. ISSN 0092-8674, 1097-4172. doi:10.1016/j.cell.2013.03.035.

[35] Xiong Ji, Daniel B. Dadon, Benjamin E. Powell, Zi Peng Fan, Diego Borges-Rivera, Sigal Shachar, Abraham S. Weintraub, Denes Hnisz, Gianluca Pegoraro, Tong Ihn Lee, Tom Misteli, Rudolf Jaenisch, and Richard A. Young. 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell*, 18(2):262–275, February 2016. ISSN 1934-5909, 1875-9777. doi:10.1016/j.stem.2015.11.007.

[36] Bo Li and Colin N. Dewey. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, August 2011. ISSN 1471-2105. doi:10.1186/1471-2105-12-323.

[37] Fazle E Faisal, Lei Meng, Joseph Crawford, and Tijana Milenković. The post-genomic era of biological network alignment. *EURASIP Journal on Bioinformatics and Systems Biology*, 2015(1), December 2015. ISSN 1687-4153. doi:10.1186/s13637-015-0022-9.

[38] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, June 2006. ISSN 0027-8424. doi:10.1073/pnas.0601602103.

[39] R. Lambiotte, J.-C. Delvenne, and M. Barahona. Laplacian Dynamics and Multiscale Modular Structure in Networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76–90, July 2014. ISSN 2327-4697. doi:10.1109/TNSE.2015.2391998.

[40] Heidi K. Norton, Harvey Huang, Daniel J. Emerson, Jesi Kim, Shi Gu, Danielle S. Bassett, and Jennifer E. Phillips-Cremins. Detecting hierarchical 3-D genome domain reconfiguration with network modularity. *bioRxiv*, page 089011, November 2016. doi:10.1101/089011.

[41] Chenchen Zou, Yuping Zhang, and Zhengqing Ouyang. HSA: Integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome Biology*, 17, March 2016. ISSN 1474-7596. doi:10.1186/s13059-016-0896-1.

[42] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 0028-0836. doi:10.1038/nature11247.

[43] Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, August 2015. ISSN 1087-0156. doi:10.1038/nbt.3300.

[44] David R. Kelley, Jasper Snoek, and John Rinn. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, page gr.200535.115, May 2016. ISSN 1088-9051, 1549-5469. doi:10.1101/gr.200535.115.

[45] pytorch. PyTorch. http://pytorch.org/, 2017.

[46] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, January 2001. ISSN 1362-4962.

[47] docker docker. Docker - Build, Ship, and Run Any App, Anywhere. https://www.docker.com/, 2017.

[48] python python. Welcome to Python.org. https://www.python.org/, 2017.

[49] plotly plotly. Visualize Data, Together. https://plot.ly/, 2017.

[50] Jupyter Jupyter. Project Jupyter. http://www.jupyter.org, 2017.

[51] HiGlass: Web-based Visual Comparison And Exploration Of Genome Interaction Maps | bioRxiv. http://www.biorxiv.org/content/early/2017/03/31/121889.

[52] unreal Unreal. What is Unreal Engine 4. https://www.unrealengine.com/what-is-unreal-engine-4, 2017.

[53] Satomi Mitsuhashi, So Nakagawa, Mahoko Ueda, Tadashi Imanishi, and Hiroaki Mitsuhashi. Nanopore-based single molecule sequencing of the D4Z4 array responsible for facioscapulohumeral muscular dystrophy. *bioRxiv*, page 157040, June 2017. doi:10.1101/157040.