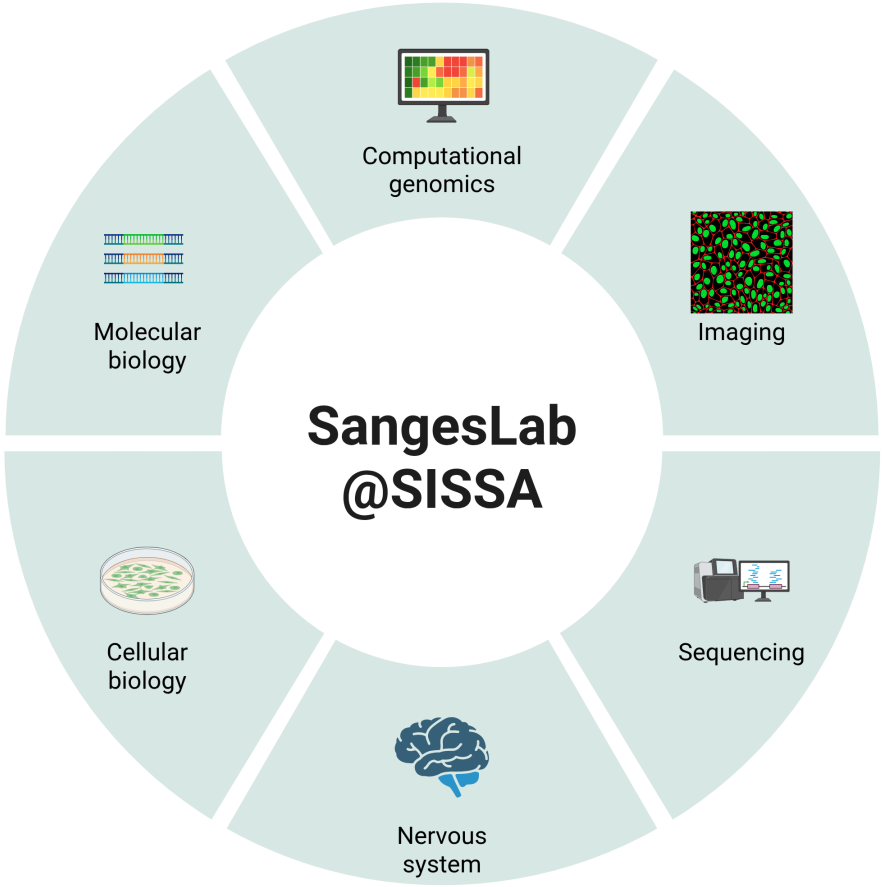
SISSA

Evolution of complexity, the non-coding genome and brain development



Thesis opportunities in Computational and Cellular Genomics

Remo Sanges

SISSA – Scuola Internazionale Superiore di Studi Avanzati
remo.sanges@sissa.it
SISSA.IT

ISTITUTO ITALIANO DI TECNOLOGIA

FANTOM6 FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

The beginning of genomics

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

The sequence of the human genome is of interest in several respects. It is the largest genome to be extensively sequenced so far, being 25 times as large as any previously sequenced genome and eight times as large as the sum of all such genomes. It is the first vertebrate genome to be extensively sequenced. And, uniquely, it is the genome of our own species.

Much work remains to be done to produce a complete finished sequence, but the vast trove of information that has become available through this collaborative effort allows a global perspective on the human genome. Although the details will change as the sequence is finished, many points are already clear.

- The genomic landscape shows marked variation in the distribution of a number of features, including genes, transposable elements, GC content, CpG islands and recombination rate. This gives us important clues about function. For example, the developmentally important HOX gene clusters are the most repeat-poor regions of the human genome, probably reflecting the very complex coordinate regulation of the genes in the clusters.
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.
- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.
- The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.
- Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark G-bands' in karyotypes.
- Recombination rates tend to be much higher in distal regions (around 20 megabases (Mb)) of chromosomes and on shorter chromosome arms in general, in a pattern that promotes the occurrence of at least one crossover per chromosome arm in each meiosis.
- More than 1.4 million single nucleotide polymorphisms (SNPs) in the human genome have been identified. This collection should allow the initiation of genome-wide linkage disequilibrium mapping of the genes in the human population.

In this paper, we start by presenting background information on the project and describing the generation, assembly and evaluation of the draft genome sequence. We then focus on an initial analysis of the sequence itself: the broad chromosomal landscape; the repeat elements and the rich palaeontological record of evolutionary and biological processes that they provide; the human genes and proteins and their differences and similarities with those of other

Lander et al. 2001

● There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.

● Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.

The beginning of genomics

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

*A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

The sequence of the human genome is of interest in several respects. It is the largest genome to be extensively sequenced so far, being 25 times as large as any previously sequenced genome and eight times as large as the sum of all such genomes. It is the first vertebrate genome to be extensively sequenced. And, uniquely, it is the genome of our own species.

Much work remains to be done to produce a complete finished sequence, but the vast trove of information that has become available through this collaborative effort allows a global perspective on the human genome. Although the details will change as the sequence is finished, many points are already clear.

- The genomic landscape shows marked variation in the distribution of a number of features, including genes, transposable elements, GC content, CpG islands and recombination rate. This gives us important clues about function. For example, the developmentally important HOX gene clusters are the most repeat-poor regions of the human genome, probably reflecting the very complex coordinate regulation of the genes in the clusters.
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.
- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.
- The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.
- Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark G-bands' in karyotypes.
- Recombination rates tend to be much higher in distal regions (around 20 megabases (Mb)) of chromosomes and on shorter chromosome arms in general, in a pattern that promotes the occurrence of at least one crossover per chromosome arm in each meiosis.
- More than 1.4 million single nucleotide polymorphisms (SNPs) in the human genome have been identified. This collection should allow the initiation of genome-wide linkage disequilibrium mapping of the genes in the human population.

In this paper, we start by presenting background information on the project and describing the generation, assembly and evaluation of the draft genome sequence. We then focus on an initial analysis of the sequence itself: the broad chromosomal landscape; the repeat elements and the rich palaeontological record of evolutionary and biological processes that they provide; the human genes and proteins and their differences and similarities with those of other

Lander et al. 2001

● There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.

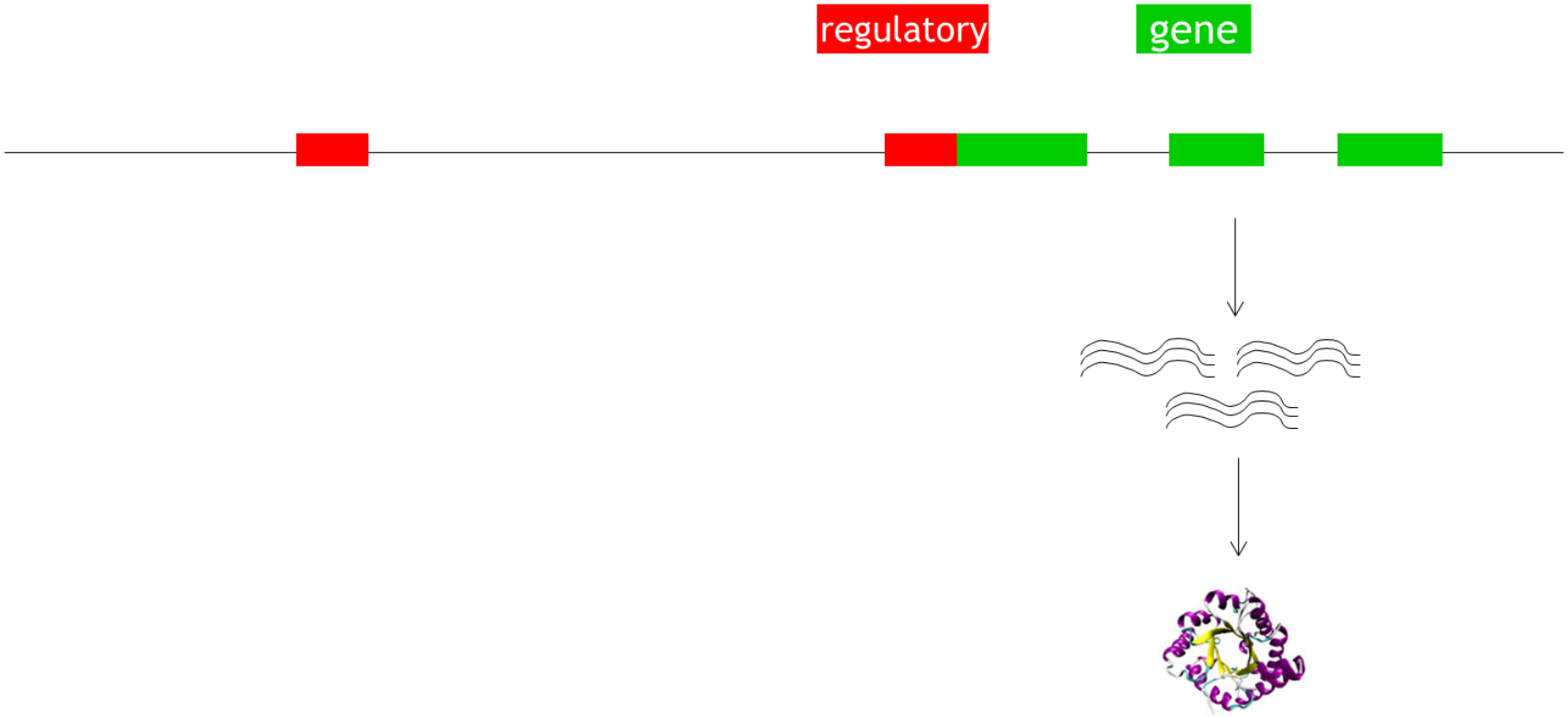
● Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.

Protein coding transcripts represent about the 3% of the human genome.

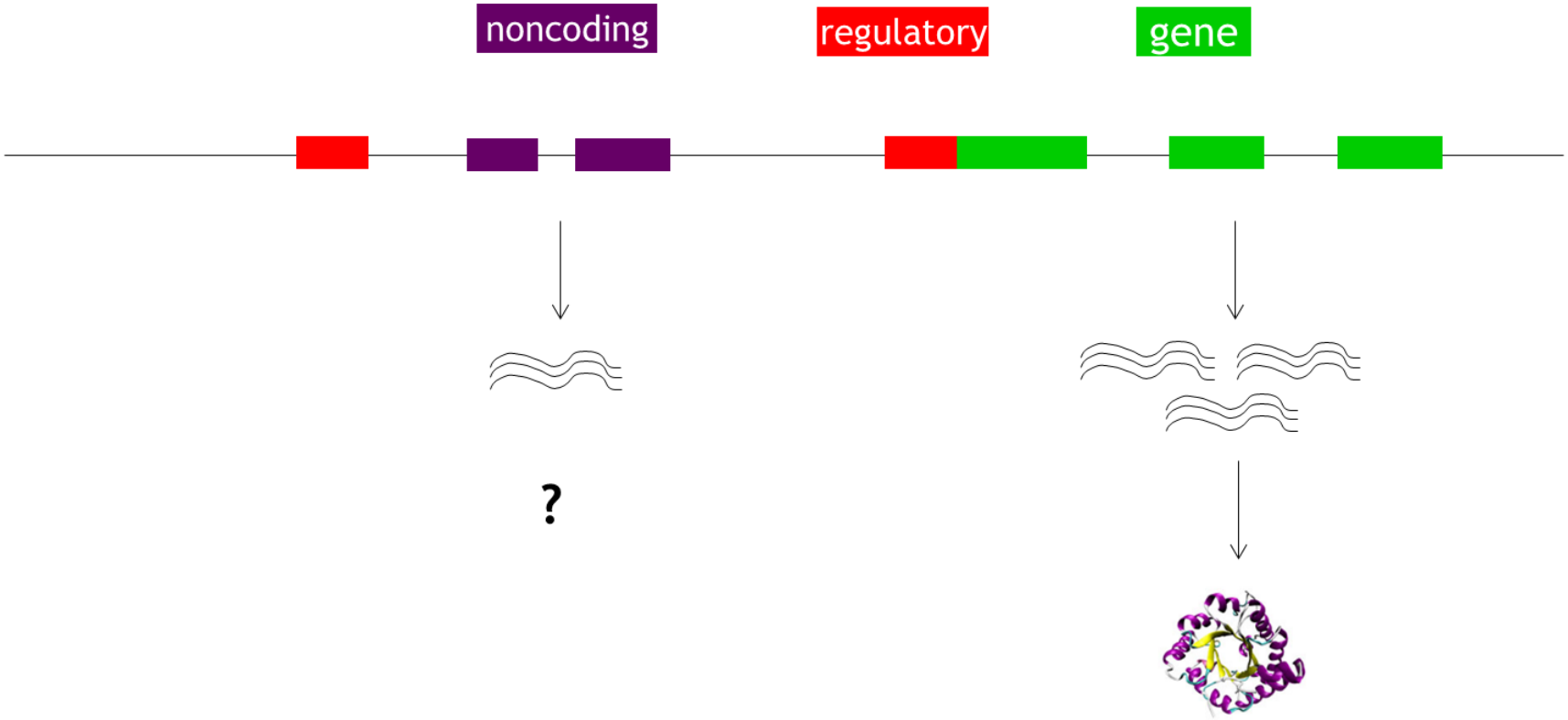
What is the function of all the remaining portion?

The Central Dogma of molecular biology was not enough

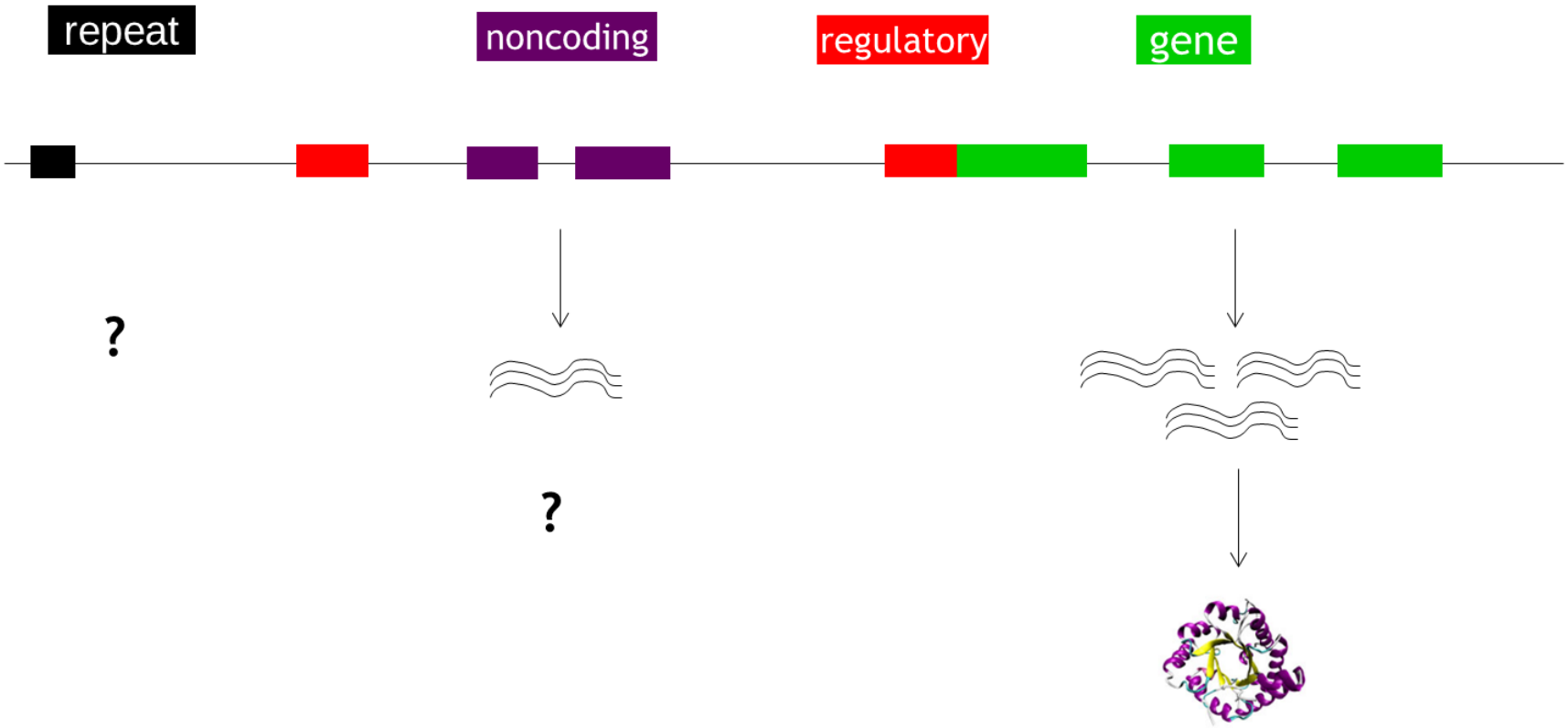
The complexity of transcription and its regulation



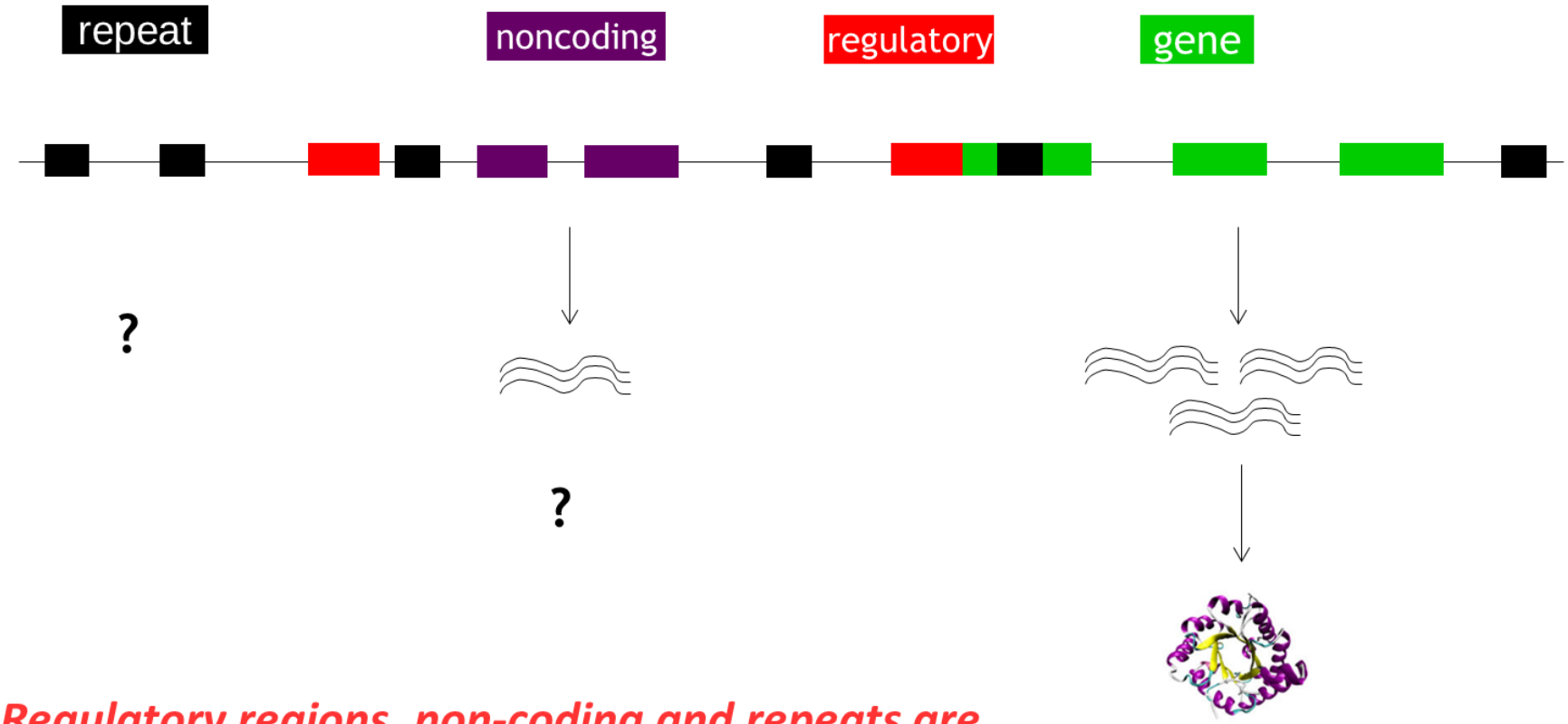
The complexity of transcription and its regulation



The complexity of transcription and its regulation

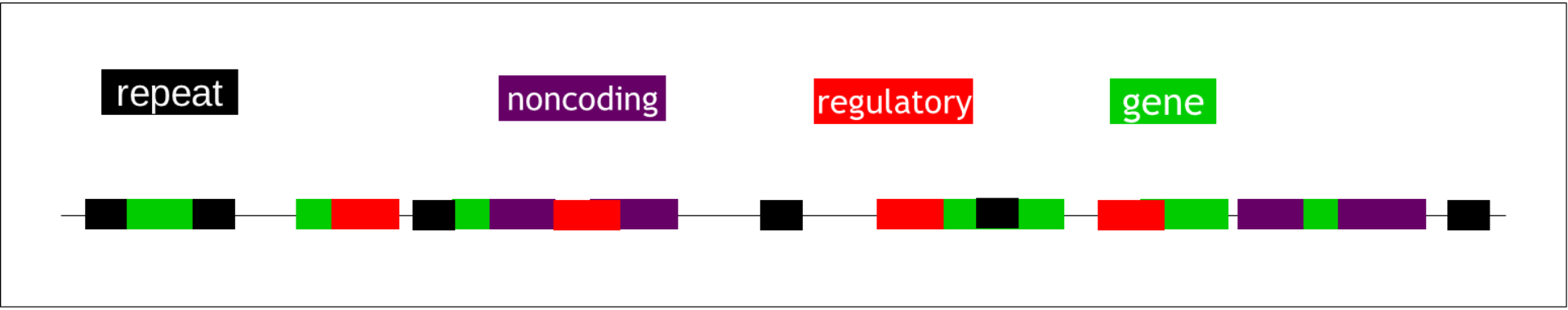


The complexity of transcription and its regulation



Regulatory regions, non-coding and repeats are considered the Dark Matter of the Genome

The complexity of transcription and its regulation



Which ones are functional?

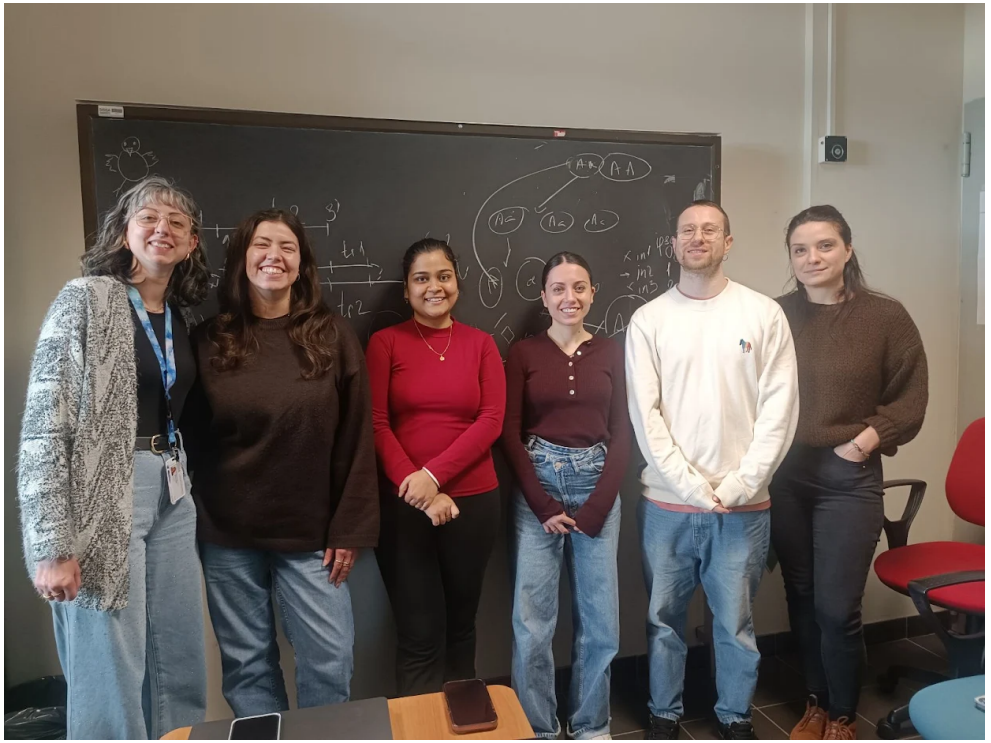
What are their functions?

How can we identify them?

How do they define complexity?

Who we are

- Computational Genomics Lab @ SISSA
- Wet lab + computational integration
- Focus on evolution, development and degeneration of the brain

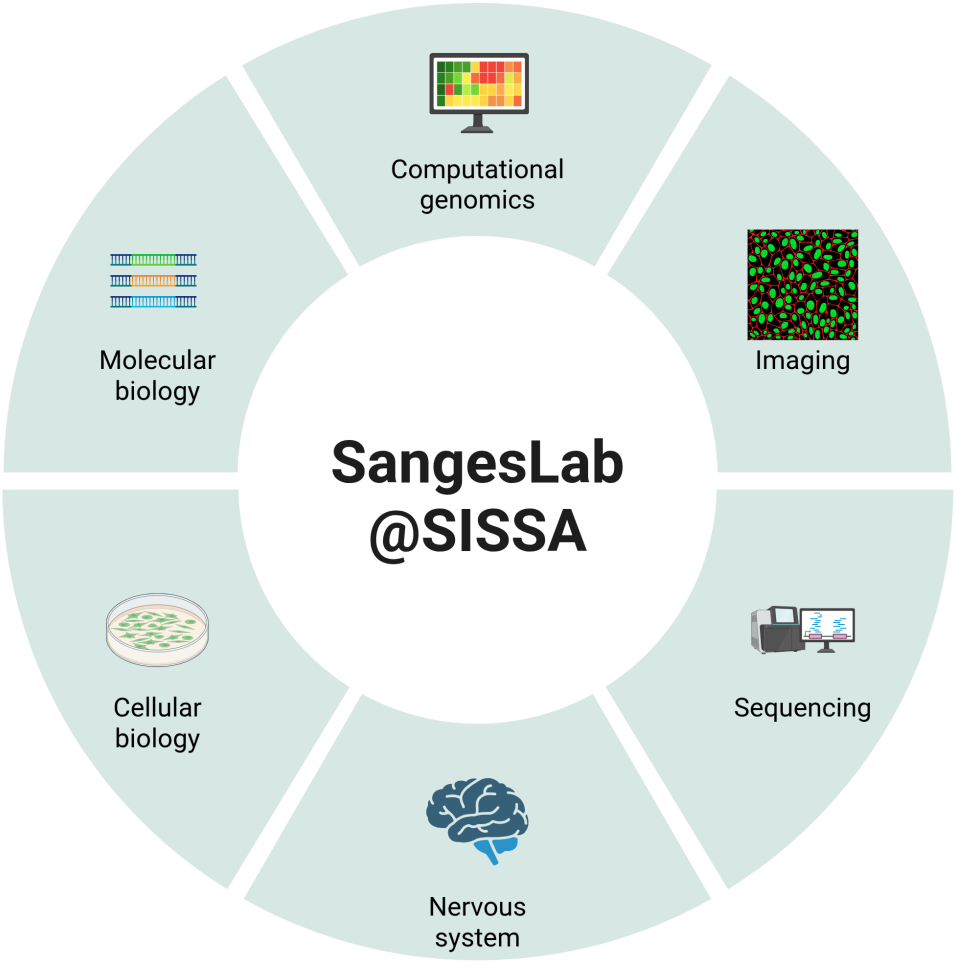


What we study

- Non-coding DNA (regulatory regions, non-coding RNAs and transposable elements)
- Chromatin dynamics and transcription factors activities
- Strategies to modulate gene expression

How we work

- Genomics data (RNA-seq, ATAC-seq, scRNA-seq...)
- Computational analysis
- Experimental validation



From data to hypothesis

- Identify candidate genes / regulatory regions
- Compare across conditions / cell types
- Generate testable hypotheses

From hypothesis to function

- Perturb the system (e.g. siRNA, shRNA, CRISPR)
- Measure gene expression changes, cellular phenotypes
- Link to biological effect

One thing we study

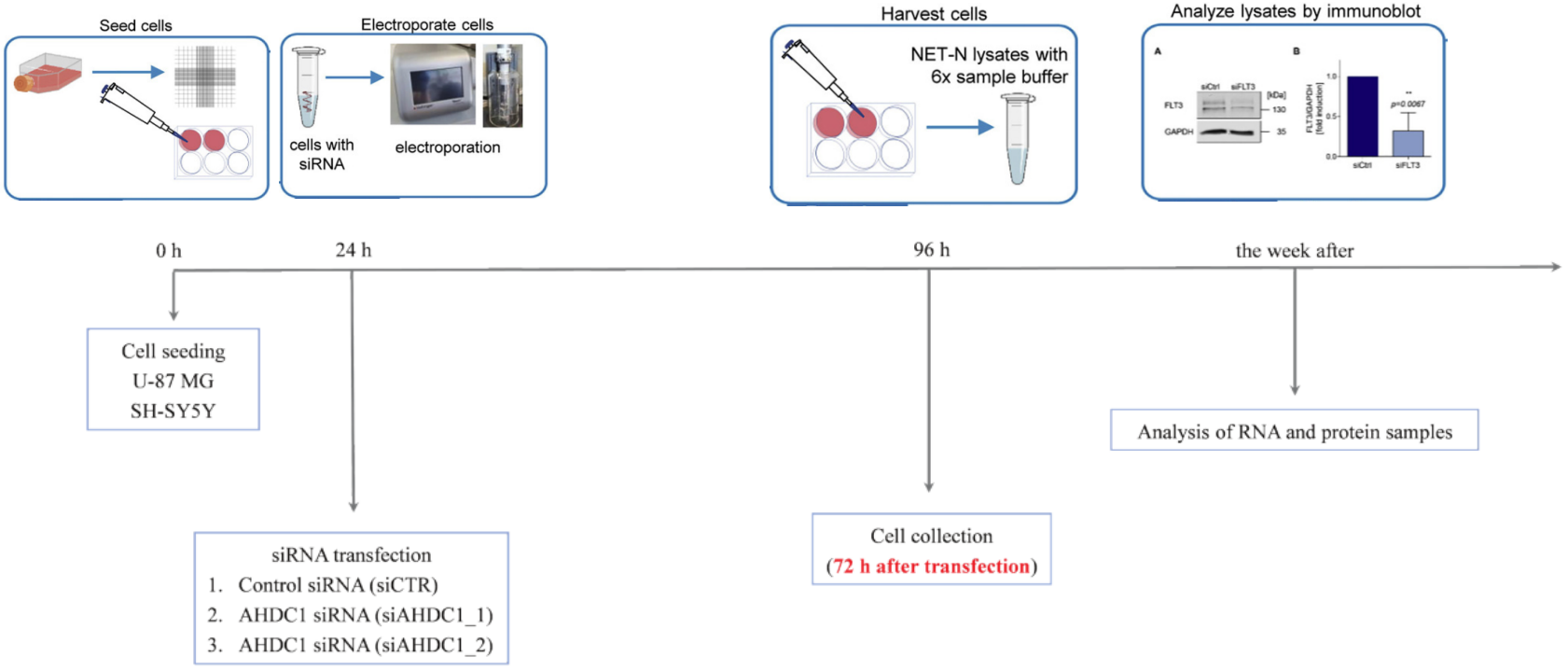
REPORT

De Novo Truncating Mutations in *AHDC1* in Individuals with Syndromic Expressive Language Delay, Hypotonia, and Sleep Apnea

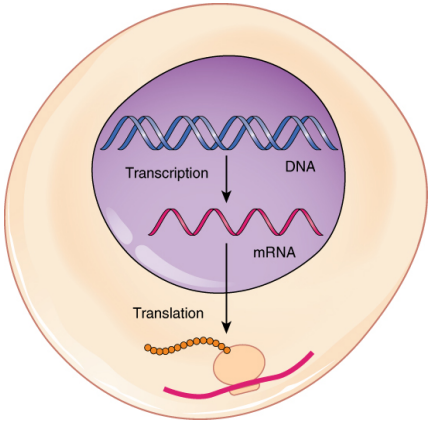
Fan Xia,[1] Matthew N. Bainbridge,[2] Tiong Yang Tan,[3,4] Michael F. Wangler,[1,5] Angela E. Scheuerle,[6] Elaine H. Zackai,[7] Margaret H. Harr,[7] V. Reid Sutton,[1,5] Roopa L. Nalam,[2,8] Wenmiao Zhu,[1] Margot Nash,[3] Monique M. Ryan,[3] Joy Yaplito-Lee,[3] Jill V. Hunter,[5] Matthew A. Deardorff,[7] Samantha J. Penney,[1] Arthur L. Beaudet,[1] Sharon E. Plon,[1,5] Eric A. Boerwinkle,[2,9] James R. Lupski,[1,5] Christine M. Eng,[1] Donna M. Muzny,[2] Yaping Yang,[1] and Richard A. Gibbs[1,2,*]

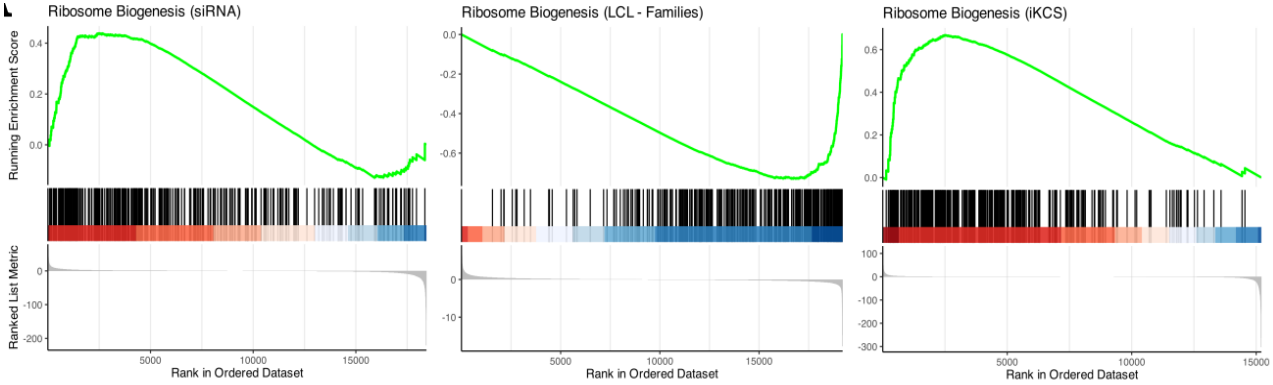
Clinical whole-exome sequencing (WES) for identification of mutations leading to Mendelian disease has been offered to the medical community since 2011. Clinically undiagnosed neurological disorders are the most frequent basis for test referral, and currently, approximately 25% of such cases are diagnosed at the molecular level. To date, there are approximately 4,000 "known" disease-associated loci, and many are associated with striking dysmorphic features, making genotype-phenotype correlations relatively straightforward. A significant fraction of cases, however, lack characteristic dysmorphism or clinical pathognomonic traits and are dependent upon molecular tests for definitive diagnoses. Further, many molecular diagnoses are guided by recent gene-disease association discoveries. Hence, there is a critical interplay between clinical testing and research leading to gene-disease association discovery. Here, we describe four probands, all of whom presented with hypotonia, intellectual disability, global developmental delay, and mildly dysmorphic facial features. Three of the four also had sleep apnea. Each was a simplex case without a remarkable family history. Using WES, we identified *AHDC1* de novo truncating mutations that most likely cause this genetic syndrome.

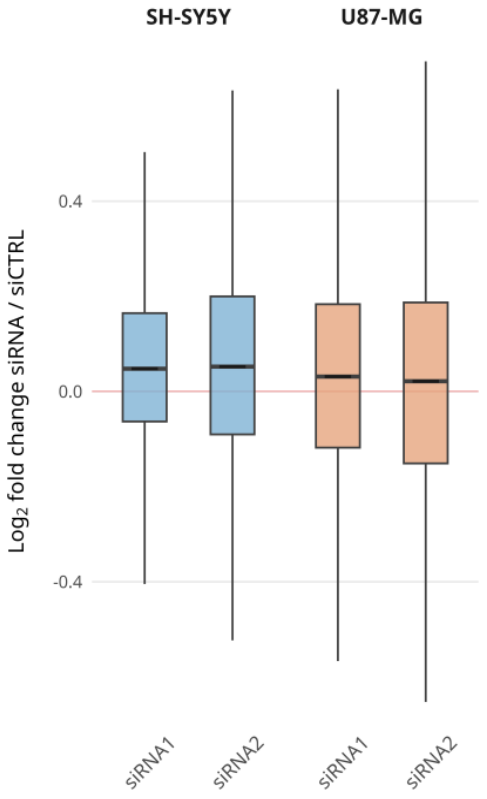
From hypothesis to function



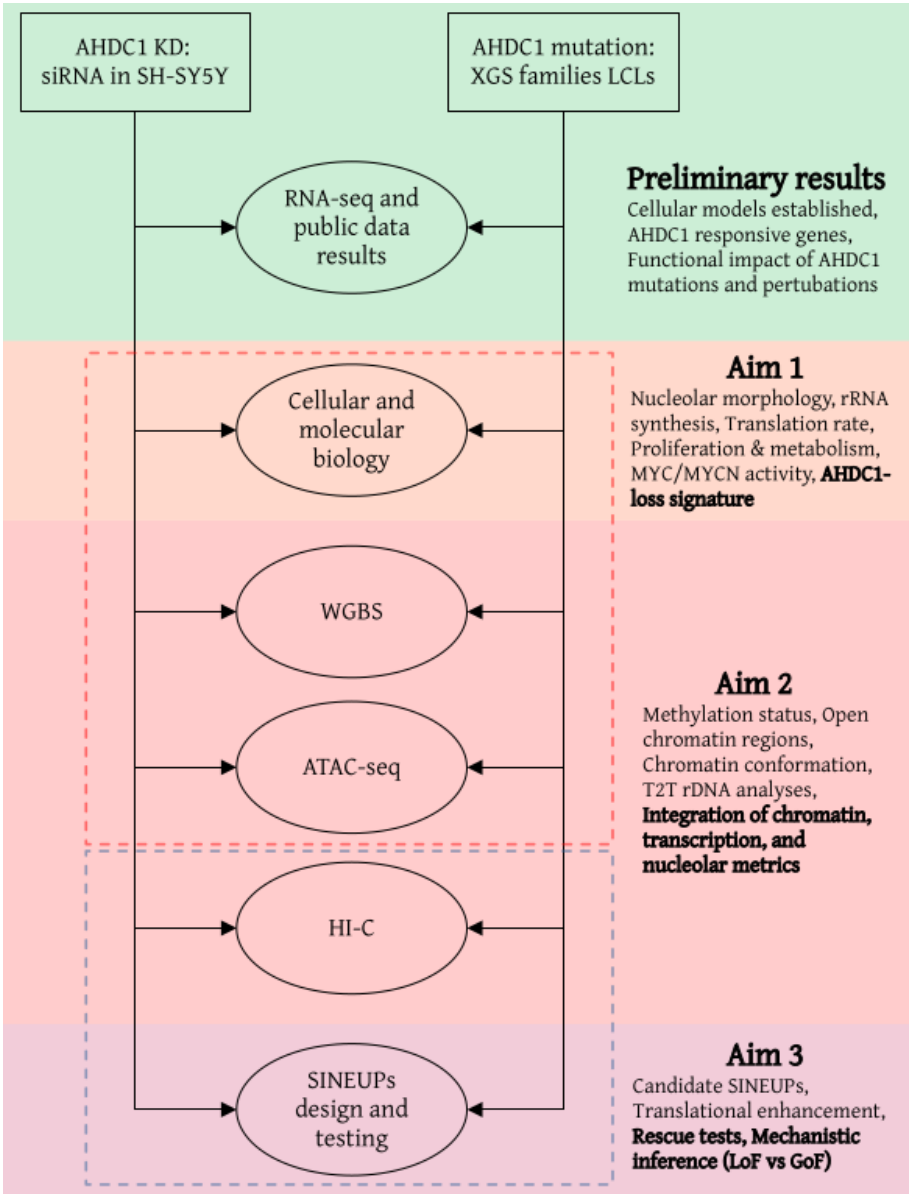
From data to hypothesis







And the cycle restart



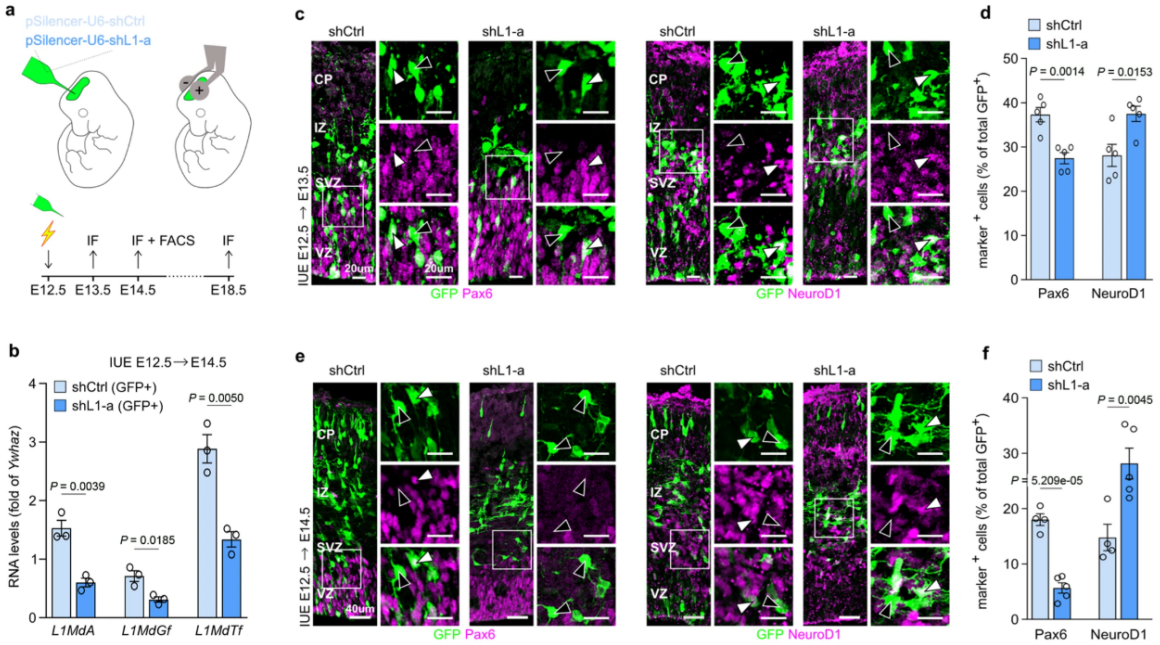
What you will learn

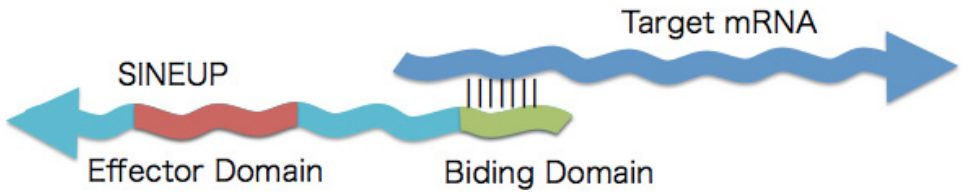
- How to interpret genomic data
- How to design experiments
- Critical thinking in biology
- Basic data analysis skills

Other current research

Fig. 1: L1 silencing alters neocortical development.

From: LINE-1 regulates cortical development by acting as long non-coding RNAs





ISTITUTO ITALIANO DI TECNOLOGIA

FANTOM6
FUNCTIONAL ANNOTATION OF THE MAMMALIAN GENOME

Interested?

remo. sanges@gmail.com