Genome structure, analysis and evolution

Lecture 1

Nunzio D’Agostino
nunzio.dagostino@entecra.it; nunzio.dagostino@gmail.com
The term **GENOME** was coined in 1920 by the German botanist **Hans Winker**

\[ \text{GENe +chromosOME} \]

"I propose the expression **GENOME** for the haploid chromosome set, which [...] specifies the material foundations of the species”

Actual definition: Entire genetic complement of an organism

The term **GENOMICS** was coined in 1986 by the geneticist **Thomas Roderick**

”mapping, sequencing and characterizing genomes”
Genome sequence can reveal us...

- everything about organism's life
- its developmental program
- disease resistance or susceptibility
- how do we struggle, survive and die
- where are we going and where we came from
- how similar are we to other organisms
structural genomics

Structural genomics allows:
- the nucleotide sequence of a genome (the DNA primary structure) to be determined
- the architecture, the organization and the complexity of the genomes to be described
Why genome sequencing

Genome sequencing provides:

Access to the complete gene catalogue for a species, to the regulatory elements that control their function and a framework for understanding genomic variation;

Genome sequence is a prerequisite resource for:
- understanding fully the roles of genes in plant development and adaptation;
- exploiting the natural and induced genetic diversity of an organism

Finally, Genome sequencing offers a new powerful tool for molecular breeding in crop plants.
DNA compaction is very complex and the DNA isn’t just crammed into the nucleus but is organized in a very orderly fashion from the smallest unit - the **nucleosome** to the entire **chromosome** which has a fixed space in the nucleus.

The importance of DNA packing:
1. To protect DNA from damage
2. DNA in a chromosome can be transmitted efficiently to both daughter cells during cell division
3. Chromosome confers an overall organization to each molecule of DNA, which facilitates gene expression as well as recombination
Proteins in chromosomes

Half of the molecular mass of eukaryotic chromosome is determined by **proteins**

- In eukaryotic cells a given region of DNA with its associated proteins is called **chromatin**
- The majority of the associated proteins are small, basic proteins called **histones**.
- Other proteins associated with the chromosome are referred to as **non-histone proteins**, including numerous DNA binding proteins that regulate the transcription, replication, repair and recombination of DNA.
Nucleosomes are the building blocks of chromosomes

octamer of core histones: H2A, H2B, H3, H4 (each one \( \times 2 \))
High order chromatin structure

Nucleosoma

- Core di 8 molecole di istori H2A, H2B, H3 e H4
- Istone H1
- DNA centrale
- DNA di collegamento

DNA a doppia elica

Collana di perle

Compacts DNA 7X more

- 30 nm

Fibra ad ansa

Impalcatura condensata

Cromosoma mitotico

DNA di collegamento
Chromatin has two forms:

- **euchromatin**
  - When stained and observed under an optical microscope, euchromatins are the light-colored bands while heterochromatins are the dark-colored bands.
  - Darker staining indicates tighter DNA packaging. Heterochromatins thus have tighter DNA packaging than euchromatins.
  - Heterochromatins are compactly coiled regions while euchromatins are loosely coiled regions.
  - Euchromatin contains less DNA while heterochromatin contains more DNA.
  - Euchromatin is early replicative while heterochromatin is late replicative.
  - Heterochromatin is transcriptionally inactive and contains highly repeated DNA sequences, such as those present at centromeres and telomeres.
Heterochromatin & euchromatin boundary is dynamic

When a gene that is normally expressed in euchromatin is experimentally relocated into a region of heterochromatin, it ceases to be expressed, and the gene is said to be silenced. These differences in gene expression are example of *position effects*, in which the activity of a gene depends on its position relative to a nearby region of heterochromatin on a chromosome.
Constitutive: It is composed of DNA sequences that are not normally ever expressed.

Facultative: It includes DNA sequences that appear to be expressed in a given moment of the cell development program or in specific cell types. The biological significance of facultative heterochromatin reflects the genetic activities of the different cell types specialized to perform different functions.
Constitutive

- It contains a particular type of DNA called satellite DNA, which consists of a large number of short tandemly repeated sequences. These satellite DNA sequences are able to fold on themselves and may have an important role in the formation of the highly compact structure of constitutive HC.

- It is stable and conserves its heterochromatic properties during all stages of development and in all tissues.

- It is highly polymorphic, probably because of the instability of the satellite DNA. This polymorphism can affect not only the size but also the localization of the heterochromatin, and apparently has no phenotypic effects.
**Facultative**

- It is characterised by the presence of LINE-type repeated sequences. These sequences, dispersed throughout the genome, could promote the propagation of a condensed chromatin structure;

- It is reversible, its heterochromatic state depends on the stage of development or on the cell type examined

- It is not particularly rich in satellite DNA, and it is therefore not polymorphic
euchromatin vs heterochromatin
Genome size & the complexity of the organism

- **Genome size**: the length of DNA associated with one haploid complement of chromosomes
- **Gene number**: the number of genes included in a genome
- **Gene density**: the average number of genes per Mb of genomic DNA
This larger size of eukaryotic genomes is not inherently surprising, since one would expect to find more genes in organisms that are more complex. However, the genome size of many eukaryotes does not appear to be related to genetic complexity. For example, the genomes of salamanders and lilies contain more than ten times the amount of DNA that is in the human genome, yet these organisms are clearly not ten times more complex than humans.

The range of sizes of the genomes of representative groups of organisms are shown on a logarithmic scale.
C-value paradox

C-value = size of the haploid genome

There is no correlation between complexity of an organism and the size of its genome
Estimating DNA C-values (genome size)

Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects
M. D. BENNETT* and I. J. LEITCH

- Feulgen microdensitometry (Fe)
- Flow cytometry (FC)
- Computer-based image analysis (CIA) systems, which can estimate DNA amounts using Feulgen-stained cytological preparations in place of a microdensitometer.
Since 2000 the scientific and popular press has reported and celebrated the ‘complete’ sequencing of the first insect (Drosophila melanogaster) and plant genome (Arabidopsis thaliana) and the human genome (in 2001). For example, a title in Nature reported: ‘The sequencing of an entire plant genome is now complete.’ Readers could be forgiven for assuming this meant the entire linear sequence of the nuclear DNA had been sequenced and assembled, so that the total size of the nuclear genome in these organisms was now known with certainty, and hence much more accurately than any previous estimate based on other methods subject to various experimental errors.

Further potential for confusion comes from new uses of the term ‘genome’ recently spawned by genome sequencers. These concern the counter-intuitive meaning of a ‘wholly’, ‘completely’ or ‘entirely’ sequenced genome, or of equating ‘genome’ with ‘euchromatic genome’—a confusing concept in which ‘genome’ equals the parts which could be cloned and sequenced, but not the rest (see below).

**Exact C-values based on complete genome sequences would be invaluable. The need to complete sequencing gaps (...) remains technically difficult, and it is unclear how, when, or if it will be achieved. Genome sequencing becomes more difficult as genome size increases, and experience with Arabidopsis implies that exact C-values are unlikely to be obtained in this way soon for any larger plant genomes.**
The Plant DNA C-values database provides a one-stop, user-friendly db where plant genome sizes can be readily accessed and compared.
Plant genomes: total size

- Human
- Cotton
- Barley
- Sugarcane
- Wheat
Plant genomes: total size

*Genlisea margaretae*

63 Mb = (0.0648 pg)

*Paris japonica*

148.852 Mb = (150.20 pg)

2400 times more
Arrangement of plant genome: an overview

- many repetitive sequences
- tandem gene duplications and duplications of long chromosomal regions
- **plants with big genomes**: genes present in „gene-rich“ islands isolated with long regions of non-coding (often repetitive) DNA.
The number of chromosomes is not correlated to the total genome size

**Oryza sativa** (rice)

Genome size ≈489 Mb,
Chr. number: n=12; 2n=24

**Hordeum vulgare** (barley)

Genome size ≈5429 Mb,
Chr. number: n=7; 2n=14
Variation in chromosome number

Genetic variability forms the basis of plant improvement and variation in chromosome number adds to genetic variability

**EUPLOID:** Chromosome number is changed to exact multiple of the basic set

**Polyploids** are euploids in multiple of basic set of chromosome
Auto- & Allo- polyploids

EUPLOIDS may be:

**AUTOPOLYPLOIDS**: having duplicate genome of same species

*Autotetraploid*: having duplicate genome of same diploid species.

**ALLOPOLYPLOIDS**: having duplicate genome of different species

*Allohexaploid*

<table>
<thead>
<tr>
<th>wheat (<em>Triticum turgidum</em>)</th>
<th>+</th>
<th>rye (<em>Secale cereale</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4x</td>
<td>+</td>
<td>2x</td>
</tr>
</tbody>
</table>

Triticale 6x

**ANEUPLOID**: Chromosome number of basic chromosome set is changed by addition or deletion of specific chromosomes
<table>
<thead>
<tr>
<th>Species</th>
<th>Crop</th>
<th>Basic Chromosome Number (x)</th>
<th>Haploid (Gametic) Number (n)</th>
<th>Somatic (Diploid) Chromosome number (2n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avena strigosa</td>
<td>Oats</td>
<td>7</td>
<td>7</td>
<td>2n = 2x= 14</td>
</tr>
<tr>
<td>Avena barbata</td>
<td>Oats</td>
<td>7</td>
<td>14</td>
<td>2n = 4x= 28</td>
</tr>
<tr>
<td>Avena sativa</td>
<td>Oats</td>
<td>7</td>
<td>21</td>
<td>2n = 6x= 42</td>
</tr>
<tr>
<td>Gossypium arboreum</td>
<td>Cotton</td>
<td>13</td>
<td>13</td>
<td>2n = 2x= 26</td>
</tr>
<tr>
<td>Gossypium hirsutum</td>
<td>Cotton</td>
<td>13</td>
<td>26</td>
<td>2n = 4x= 26</td>
</tr>
<tr>
<td>Triticum monococum</td>
<td>Wheat</td>
<td>7</td>
<td>7</td>
<td>2n = 2x= 14</td>
</tr>
<tr>
<td>Triticum turgidum</td>
<td>Wheat</td>
<td>7</td>
<td>14</td>
<td>2n = 4x= 28</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>Wheat</td>
<td>7</td>
<td>21</td>
<td>2n = 6x= 42</td>
</tr>
</tbody>
</table>
Induction of ploidy

Natural Induction may arise from

- Unreduced gametes: chromosome number is not reduced during meiosis
- Natural wide crossing following chromosome doubling

Artificial Induction:

- Environmental Shock
- Chemical
  
  Colchicine: acts by dissociating the spindle and preventing migration of the daughter chromosomes to poles
Properties of polyploids

Why is polyploidy important?

- Very common in some groups. Upwards of 70% of angiosperms (flowering plants) are polyploids--some botanists estimate 90%.

- Sometimes can be a form of “instant speciation.”

Some properties of polyploids:

- Odd-ploids (e.g., triploids) are almost always sterile.

- Even-ploids (e.g., tetraploids) are frequently fertile, but not always...

- Because they have larger nuclei, poyploid cells tend to be larger than haploid or diploid cells, and cell size (or volume) increases with ploidy level. The last property is important in two agricultural/industrial applications.
Significance of Polyploidy in Plant Breeding

- Permits greater expression of existing genetic diversity
- Helps to change the character of a plant by altering number of genomes consequently changing dosage of alleles related to particular trait
- Polyploids with uneven number of genomes (Like Triploid and Pentaploids) may result in infertility. This loss of seed production can be used to produce seedless watermelons and banana

- Most of the natural polyploids are alloploids
- All species don't exhibit vigor with increase in ploidy
- Optimum ploidy level for corn is diploid as compared to tetraploid
- Optimum ploidy level of banana is triploid (Seedless)
- Blackberry is insensitive to ploidy level
Functional genomics allows:
- gene function to be identified and described
- mechanisms that regulate gene expression and that determine interactions between genes to be outlined
The ENCODE project goal is to identify all the functional elements of genomic DNA, in both coding and non-coding regions.

The pilot phase of the Project is focused on a specified 30 megabases (~1%) of the human genome sequence and is organized as an international consortium of computational and laboratory-based scientists working to develop and apply high-throughput approaches for detecting all sequence elements that confer biological function. The results of this pilot phase will guide future efforts to analyze the entire human genome.
Produced with exclusive support from:

illuminaweb
What is a gene, post-ENCODE? History and updated definition

Mark B. Gerstein, Can Bruce, Joel S. Rozowsky, et al.

*Genome Res.* 2007 17: 669-681
Access the most recent version at doi:10.1101/gr.6339607