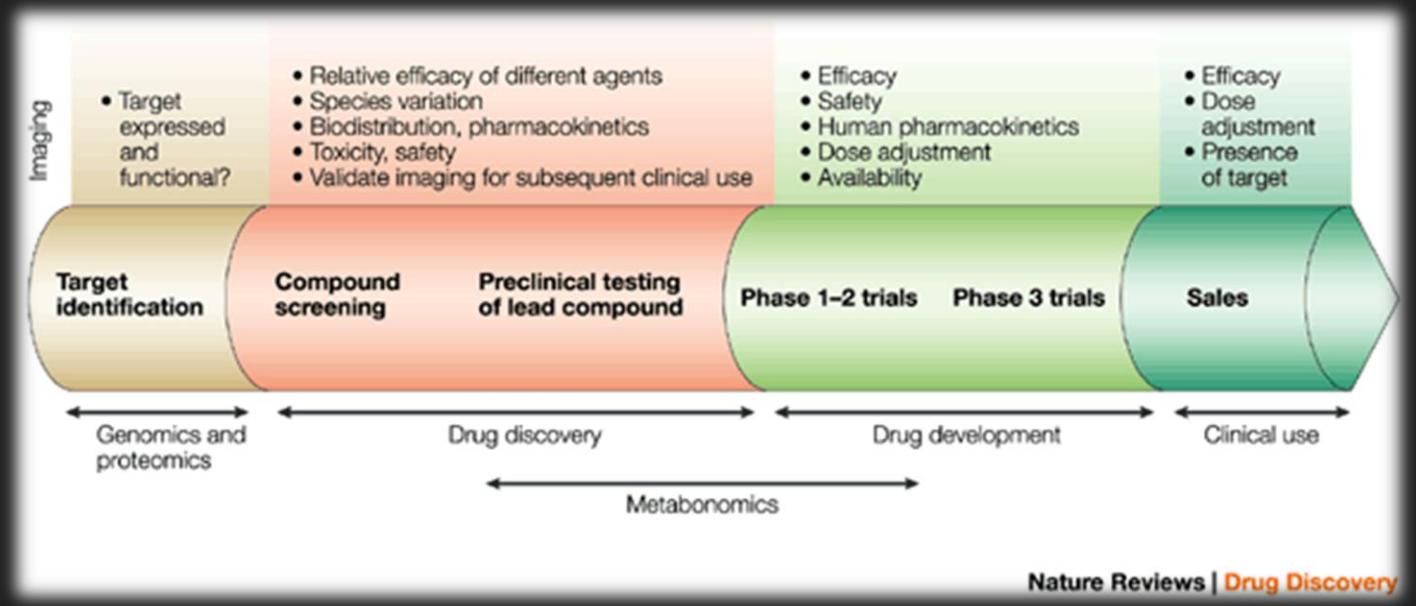
Combinar R y Python en Oncología

Amalia Martínez Segura

¿Qué hace un bioinformático en la industria farmacéutica?



- Los bioinformáticos podemos participar en todos los estadios de desarrollo de un medicamento

Dogma fundamental de la biología

ADN

- Libro de instrucciones
- Todas las células tienen la misma información

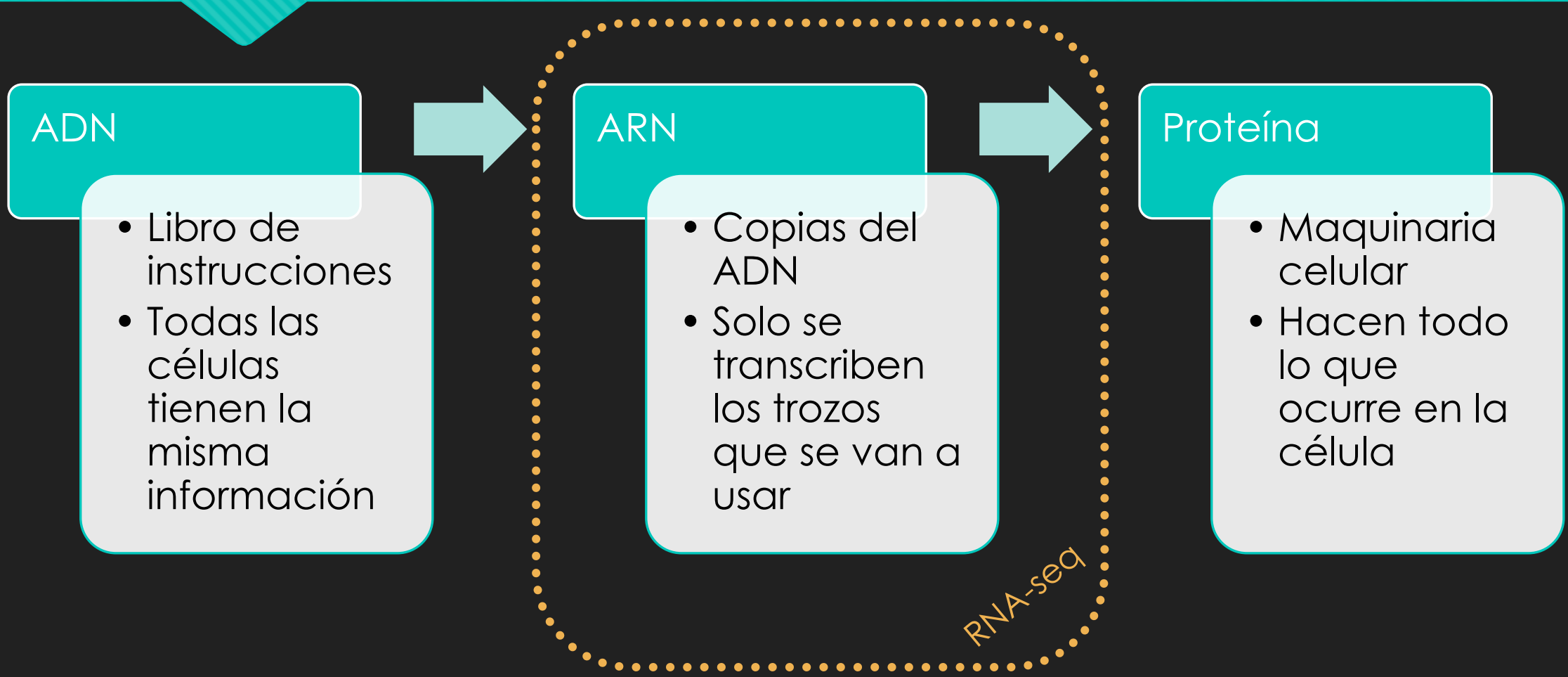
ARN

- Copias del ADN
- Solo se transcriben los trozos que se van a usar

Proteína

- Maquinaria cellular
- Formadas por aminoácidos

Dogma fundamental de la biología



Datos de secuenciación masiva

1

Secuenciación

2

Datos sin procesar

Header Sequence Quality

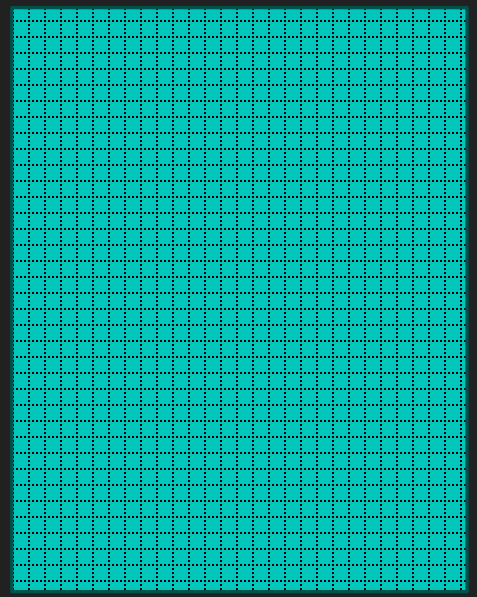
```
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHHHJIJJJJJJJJIJJJIGIGIGGIJJIJIJJJJJJIII
@HWI-ST227:389:C4WA2ACXX:7:1205:15214:42893
GAGGATCCCAGGGAGGAAGGTCCTCGCTCCTCTTTCATCTAAGGGA
+
12BAFB?A:3<AE1@<FF;1*@EG*)?0?DBD>9BF9B*?######
@HWI-ST227:389:C4WA2ACXX:8:2208:2467:44624
AAAGAGGAGAGAGGACCATCCTCCCTGGGATCCTCAGAAGTCTACT
+
BDDA:DB?2AA@FC>F?EEGC<FED>GFD;?GBB?<?F99*/9?9?
```

3

Matriz numérica

Células

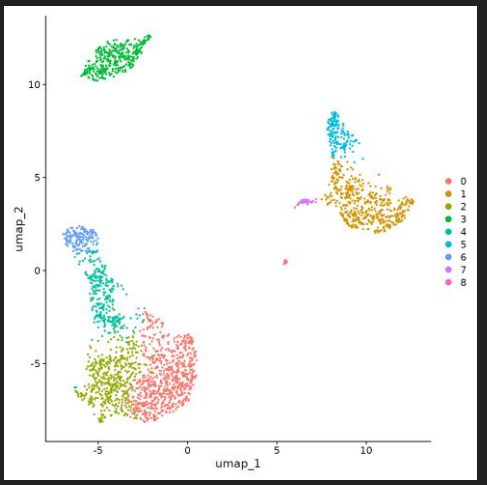
Genes



Metadata

4

Clustering y visualización



Pasos básicos del análisis

1. QC de los datos del secuenciador
2. Transformar los datos del secuenciador en una matriz numérica
3. QC de las células
4. Reducción de dimensionalidad
5. Análisis posteriores

Pasos básicos del análisis

Command line

1. QC de los datos del secuenciador
2. Transformar los datos del secuenciador en una matriz numérica
3. QC de las células
4. Reducción de dimensionalidad y visualización
5. Análisis posteriores

R o Python

Orquestrando workflows con Snakemake

- "Tool to create **reproducible and scalable** data analyses. Workflows are described via a human readable, Python based language. "

```
rule select_by_country:
        input: "data/worldcitiespop.csv"
        output: "by-country/{country}.csv"
        shell: "xsv search -s Country '{wildcards.country}' {input} > {output}"
```

```
Snakemake –c 1 by-country/Spain.csv
wildcard = 'Spain'
```

Orquestrando workflows con Snakemake

- “Tool to create **reproducible and scalable** data analyses. Workflows are described via a human readable, Python based language. ”

```
rule select_by_country:
        input: "data/worldcitiespop.csv"
        output: "by-country/{country}.csv"
        shell: "xsv search -s Country '{wildcards.country}' {input} > {output}"
```

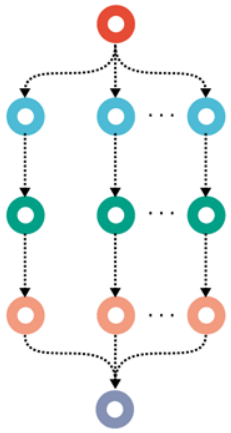
```
Snakemake –c 1 by-country/Spain.csv by-country/UnitedKingdom.csv
wildcard = [‘Spain’,’UnitedKingdom’]
```

Un workflow básico en snakemake

a

```
configfile: "config.yaml"

rule all:
    input:
        expand(
            "results/plots/{country}.hist.pdf",
            country=config["countries"]
        )

rule download_data:
    output:
        "data/worldcitiespop.csv"
    log:
        "logs/download.log"
    conda:
        "envs/curl.yaml"
    shell:
        "curl -L https://burntsushi.net/stuff/worldcitiespop.csv > {output} 2> {log}"

rule select_by_country:
    input:
        "data/worldcitiespop.csv"
    output:
        "results/by-country/{country}.csv"
    log:
        "logs/select-by-country/{country}.log"
    conda:
        "envs/xsv.yaml"
    shell:
        "xsv search -s Country '{wildcards.country}' "
        " {input} > {output} 2> {log}"
```

b



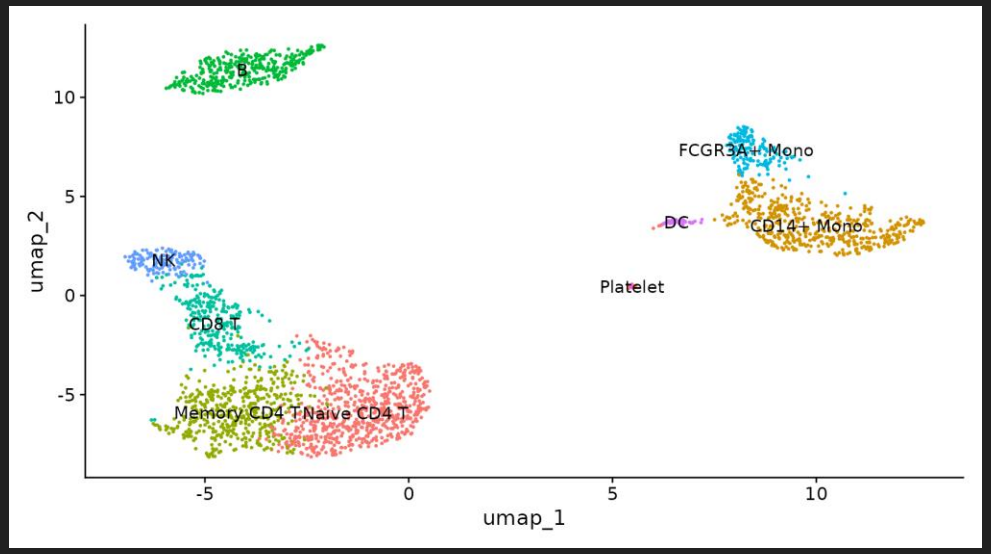
c

```
import sys
sys.stderr = open(snakemake.log[0], "w")

import matplotlib.pyplot as plt
import pandas as pd

cities = pd.read_csv(snakemake.input[0])

plt.hist(cities["Population"], bins=50)

plt.savefig(snakemake.output[0])
```

```
rule plot_histogram:
    input:
        "results/by-country/{country}.csv"
    output:
        "results/plots/{country}.hist.svg"
    container:
        "docker://faizanbashir/python-datascience:3.6"
    log:
        "logs/plot-hist/{country}.log"
    script:
        "scripts/plot-hist.py"

rule convert_to_pdf:
    input:
        "{prefix}.svg"
    output:
        "{prefix}.pdf"
    log:
        "logs/convert-to-pdf/{prefix}.log"
    wrapper:
        "0.47.0/utils/cairosvg"
```

Anotación de tipos celulares

- Anotación manual
  - Usando marcadores conocidos en la literatura
  - Sesgada



- Anotación automática
  - Menos sesgos
  - Puede que no haya referencia para los tipos celulares en tu muestra

CellTypist

Habla con tu biológo de confianza!

Otras aplicaciones

- ¿Hay algún cambio en los tipos celulares presents en la muestra?
  - Uso de herramientas de anotación automática basadas en datasets de referencia anotados manualmente
  - Pseudotiempo
- ¿Hay cambios en la expresión génica el tratamiento vs el control?
  - Análisis de expresión diferencial
- Buscar correlatos con variables de importancia clínica – por ejemplo, si el numero de linfocitos T en una muestra correlaciona con la supervivencia de un paciente de cáncer

¿Preguntas?