IRE
ISTITUTO NAZIONALE TUMORI
REGINA ELENA
ISTITUTO DI RICOVERO E CURA A CARATTERE SCIENTIFICO

Bioinformatics and Artificial Intelligence

Matteo Pallocca

UOSD Biostatistics, Bioinformatics, Clinical Trial Center

Two main blocks

1. ISAB recommendations: how we dealt with them?
2. What else happened the Data Science/Bioinfo/AI field at IRE?

2018: From the International Board report (1)

*The Bioinformatics needs **to be increased** to allow “big” data sets to be properly analyzed and include access to major data bases such as TCGA for exploratory studies*

A Brief History of NGS& Bioinformatics at IRE

Year	Field	What happened
2009	**Genomics**	Illumina GA II acquisition (first in Rome)
2013	**Bioinfo**	2 full time Bioinformaticians
2015	**Genomics**	Illumina NextSeq 500 acquisition
2016	**Genomics**	Routine Pathology Thermo Fisher NGS
2018	**Bioinfo**	3 full time Bioinformaticians
2020	**Genomics**	Illumina NovaSeq acquisition SingleCell Seque/Chromium acquisition
2020	**Bioinfo**	...

Bioinformatics Tree (Full time fellows)

Andrea Sacconi
Nanostring nCounter
miRNA/Epigenetics
Transcriptomics

Lorenzo D'Ambrosio
ACC-Immunotherapy
Immuno-Informatics

Eleonora Sperandio
CAR-T WP6.1
Machine Learning, AI, Multi-Omics

Matteo Pallocca
Coordination, Mentoring, Recruiting, Infrastructure, Projects

Stefano Scalera
Medical Oncology 2
Clinical and Pre-Clinical Genomics and Transcriptomics

Clelia Cortile
ACC-Oncohematology

Giacomo Corleone
iCARE MCurie Fellow
NGS-Epigenetics of MM

Stefano Di Giovenale
ALL-B Epigenetics

+150% Increase in Human Resources

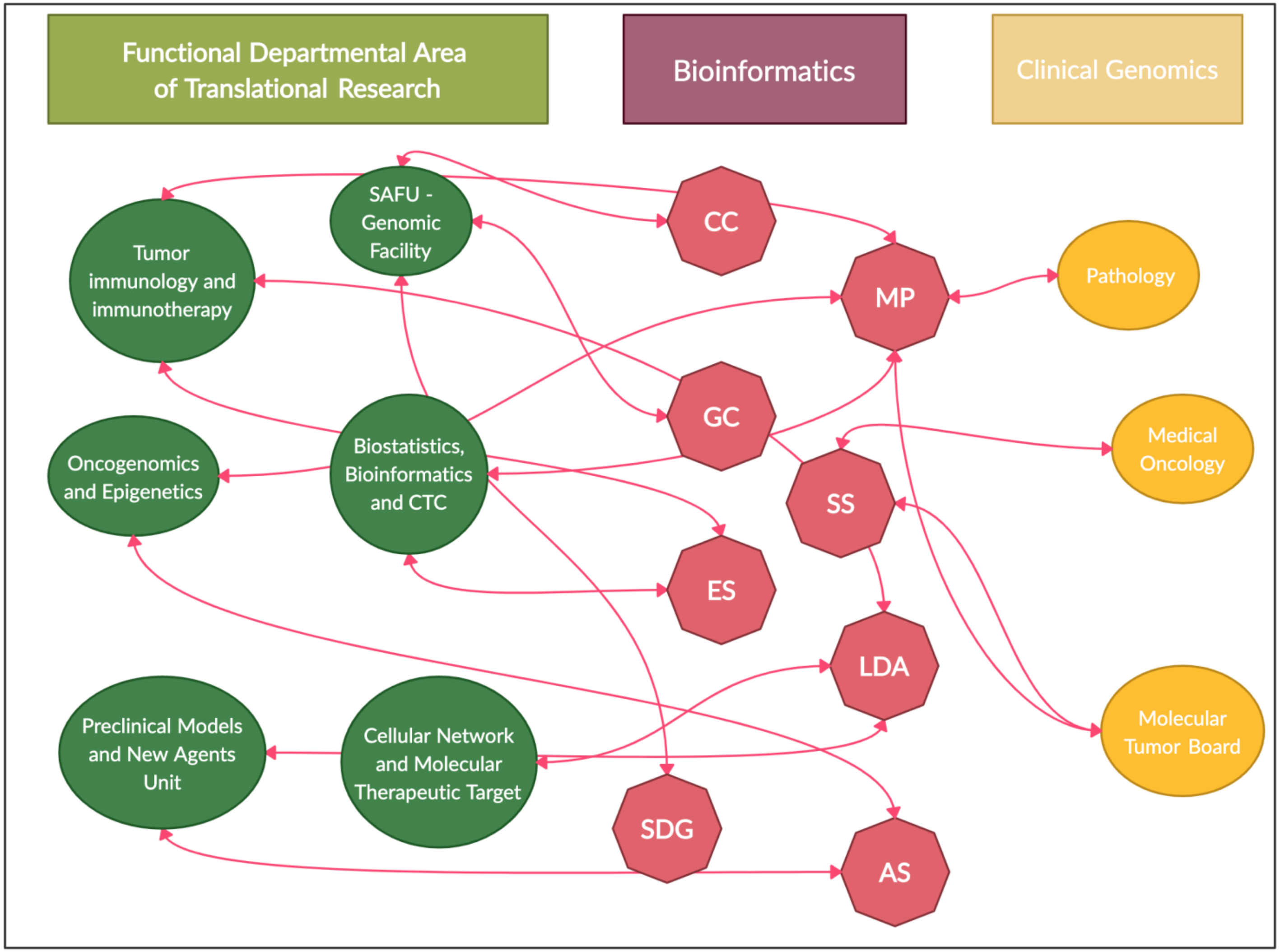
Two macro-missions carried out:

- Primary Analysis from Bio Data (90% NGS)
- Secondary Analysis, ML, AI, Visualizations

4thNov2020 ISAB

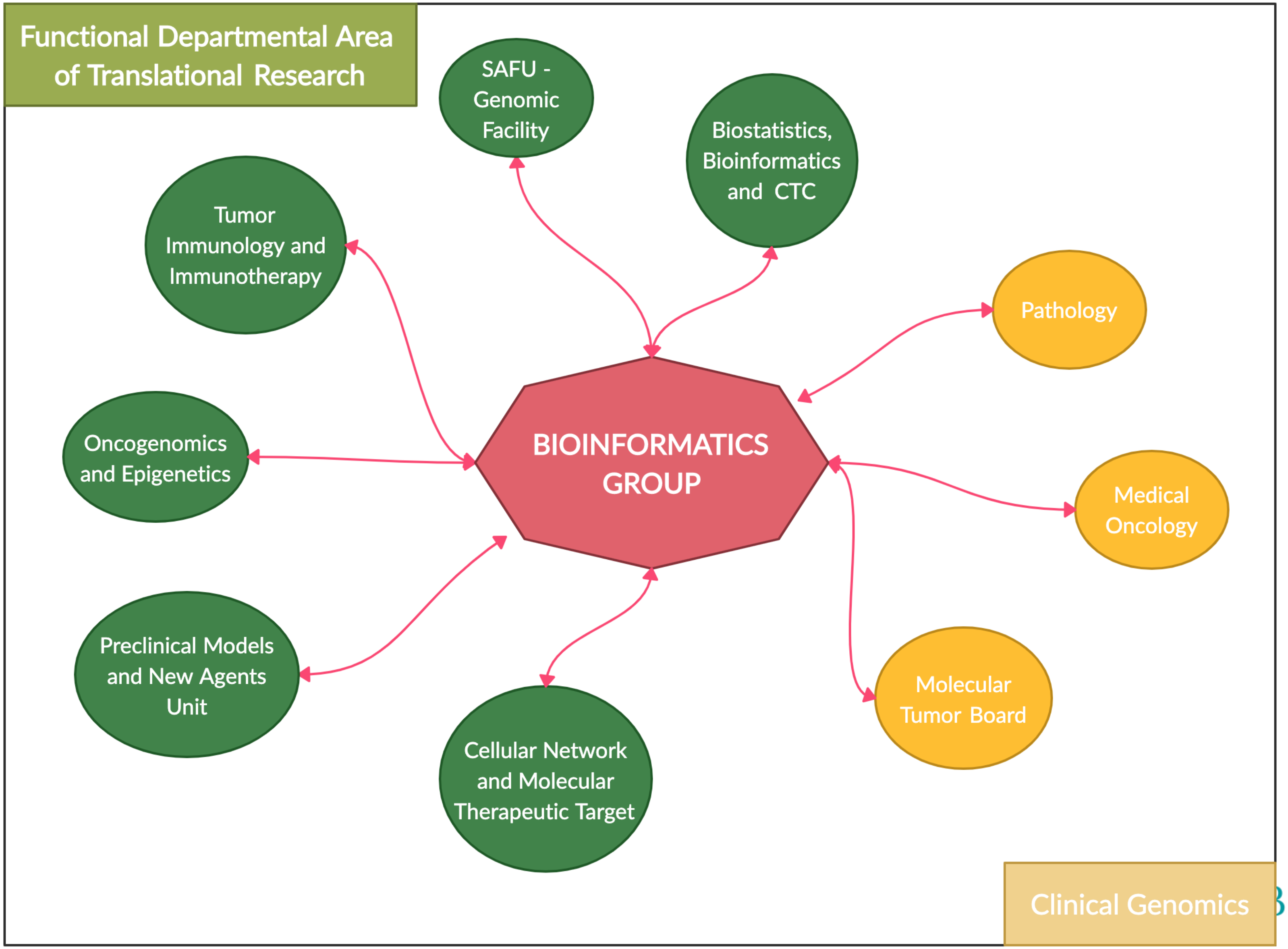
IRE

Bioinformatics Network @ IRE


Functional Departmental Area of Translational Research
Bioinformatics
Clinical Genomics
SAFU - Genomic Facility
Tumor immunology and immunotherapy
Oncogenomics and Epigenetics
Biostatistics, Bioinformatics and CTC
Preclinical Models and New Agents Unit
Cellular Network and Molecular Therapeutic Target
CC
MP
GC
SS
ES
LDA
SDG
AS
Pathology
Medical Oncology
Molecular Tumor Board

4thNov2020 ISAB IRE

Bioinformatics Network @ IRE



Functional Departmental Area of Translational Research

SAFU - Genomic Facility

Biostatistics, Bioinformatics and CTC

Tumor Immunology and Immunotherapy

Oncogenomics and Epigenetics

BIOINFORMATICS GROUP

Pathology

Medical Oncology

Preclinical Models and New Agents Unit

Cellular Network and Molecular Therapeutic Target

Molecular Tumor Board

Clinical Genomics

Which Group Model?

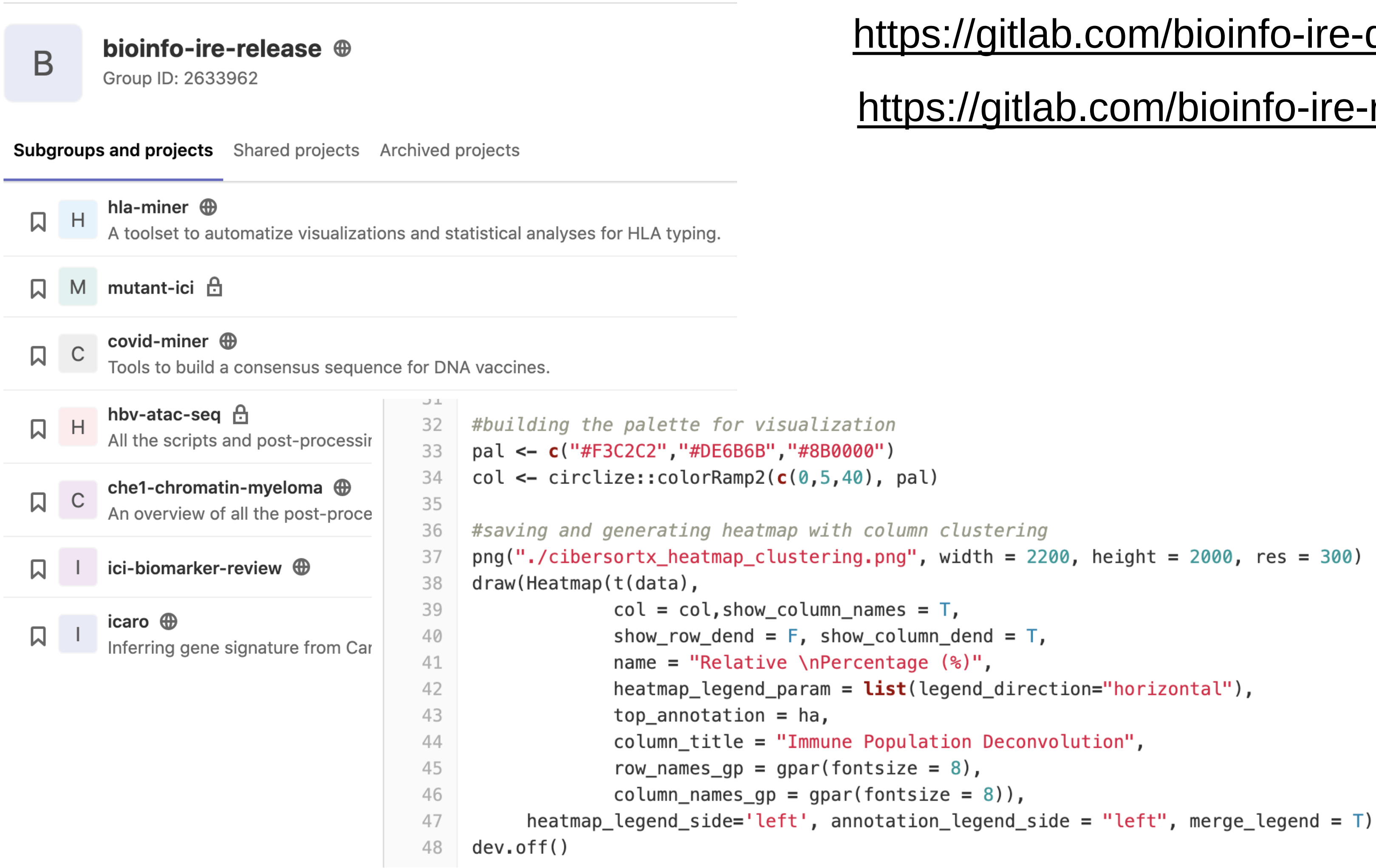
	Tutti Contro Tutti Free For All	Facility
Number of Projects	1 Each	Many
Focus / Specialization	High	Low
Process Optimization	Low	High

We employ a **hybrid model**,
attempting to have 80/20 distribution of time:
80% on personal project
20% on sharing your knowledge / supporting the facility

Challenges: Knowledge sharing (gitlab code available)

https://gitlab.com/bioinfo-ire-develop

https://gitlab.com/bioinfo-ire-release



B

bioinfo-ire-release ⊕

Group ID: 2633962

Subgroups and projects Shared projects Archived projects

- H **hla-miner** ⊕ A toolset to automatize visualizations and statistical analyses for HLA typing.
- M **mutant-ici** 🔒
- C **covid-miner** ⊕ Tools to build a consensus sequence for DNA vaccines.
- H **hbv-atac-seq** 🔒 All the scripts and post-processi
- C **che1-chromatin-myeloma** ⊕ An overview of all the post-proce
- I **ici-biomarker-review** ⊕
- I **icaro** ⊕ Inferring gene signature from Ca

```r
#building the palette for visualization
pal <- c("#F3C2C2","#DE6B6B","#8B0000")
col <- circlize::colorRamp2(c(0,5,40), pal)

#saving and generating heatmap with column clustering
png("./cibersortx_heatmap_clustering.png", width = 2200, height = 2000, res = 300)
draw(Heatmap(t(data),
             col = col,show_column_names = T,
             show_row_dend = F, show_column_dend = T,
             name = "Relative \nPercentage (%)",
             heatmap_legend_param = list(legend_direction="horizontal"),
             top_annotation = ha,
             column_title = "Immune Population Deconvolution",
             row_names_gp = gpar(fontsize = 8),
             column_names_gp = gpar(fontsize = 8)),
     heatmap_legend_side='left', annotation_legend_side = "left", merge_legend = T)
dev.off()
```

No need to *reinvent the wheel*

4thNov2020

Challenges: Knowledge sharing (gitlab code available)

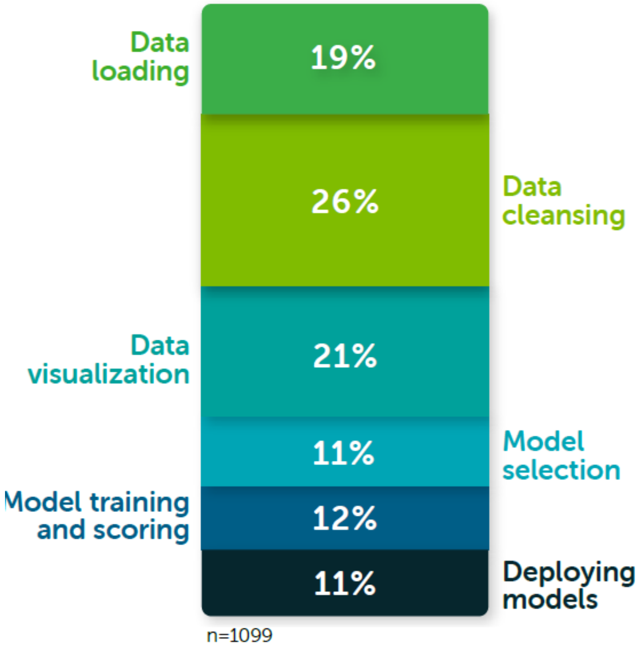
First challenge
Standardize all Primary Analyses and Primary Visualization

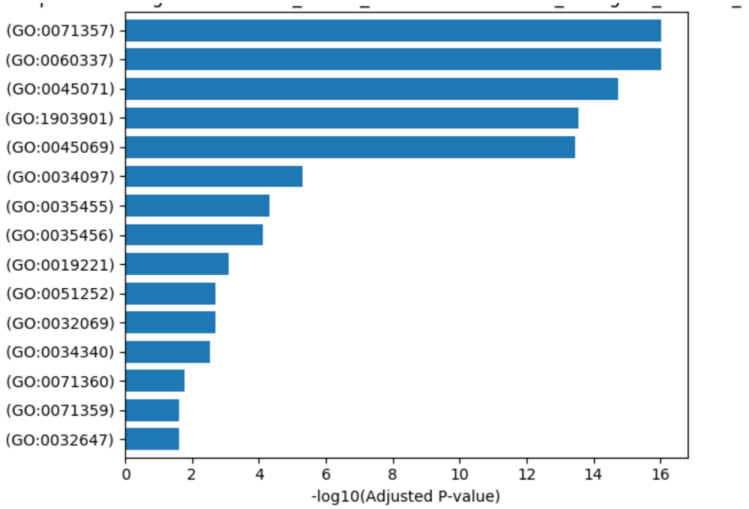
One of the most time consuming tasks in Data Science

Projects ongoing:

Auto-Go
Standardize & Automatize Gene Ontology Analyses

HLA-Miner
Standardize & Automatize HLA Haplotyping, Evolutionary Divergence development and Differential Analysis





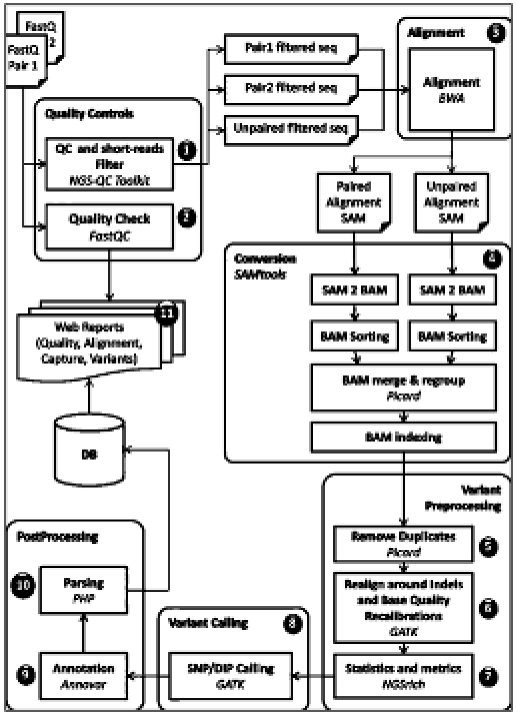
2018: From the International Board report (2)

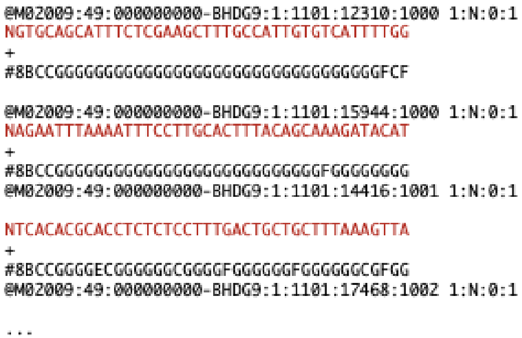
The Bioinformatics needs to be increased ***to allow “big” data sets to be properly analyzed*** *and include access to major data bases such as TCGA for exploratory studies*

Big Datasets and Big Data for Real

1 Patient with Whole Exome Sequencing:
12-24h Computing Time; >200 core-hours (parallel); 20-30GB RAM;
10-100GB storage required;

1 Patient with RNA-seq:
12-24h Computing Time; >200 core-hours (parallel); 20-30GB RAM
10-100GB storage required;

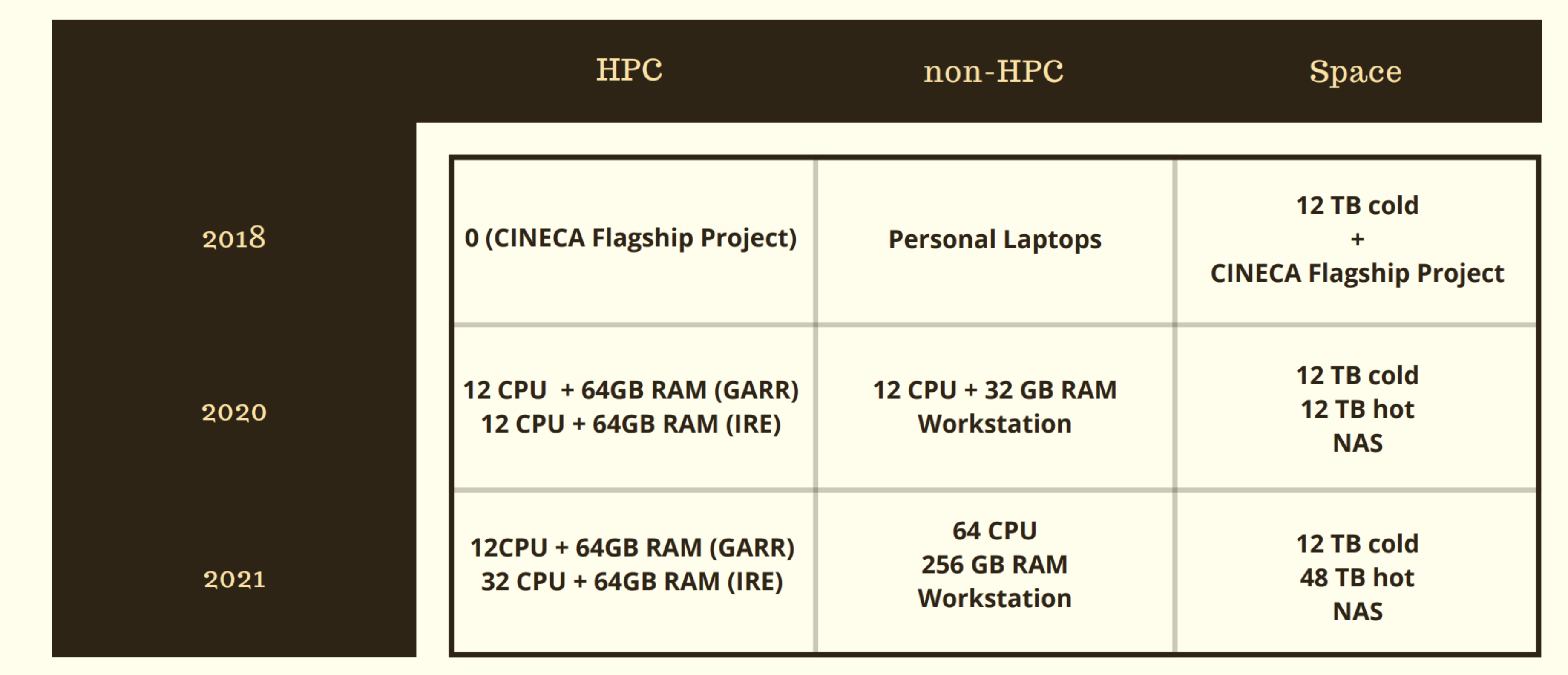




```
@M02009:49:000000000-BHDG9:1:1101:12310:1000 1:N:0:1
NGTGCAGCATTTCTCGAAGCTTTGCCATTGTGTCATTTTGG
+
#8BCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFCF

@M02009:49:000000000-BHDG9:1:1101:15944:1000 1:N:0:1
NAGAATTTAAAATTTCCTTGCACTTTACAGCAAAGATACAT
+
#8BCCGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGG
@M02009:49:000000000-BHDG9:1:1101:14416:1001 1:N:0:1

NTCACACGCACCTCTCTCCTTTGACTGCTGCTTTAAAGTTA
+
#8BCCGGGGECGGGGGGCGGGGFGGGGGGFGGGGGGCGFGG
@M02009:49:000000000-BHDG9:1:1101:17468:1002 1:N:0:1

...
```

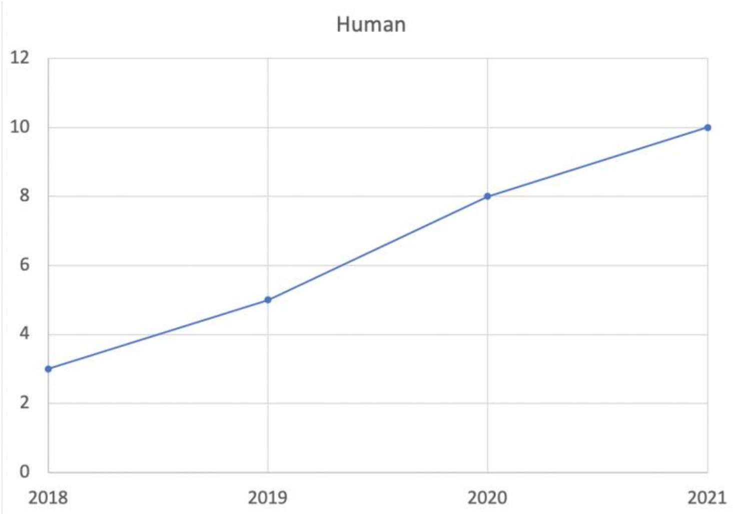
4thNov2020

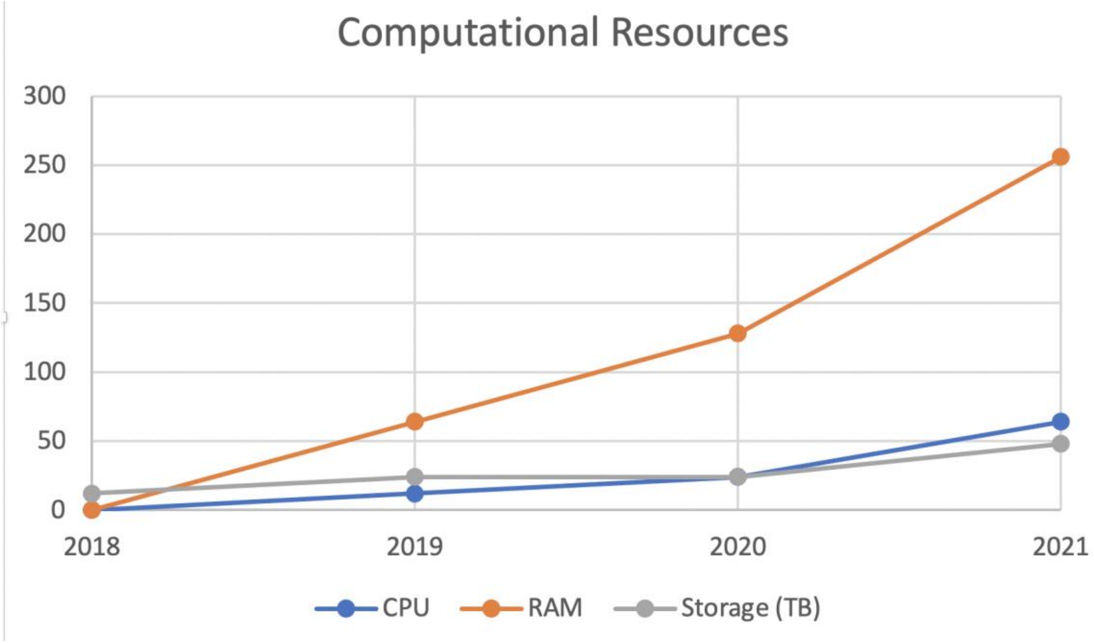
Space and time resources



	HPC	non-HPC	Space
2018	0 (CINECA Flagship Project)	Personal Laptops	12 TB cold + CINECA Flagship Project
2020	12 CPU + 64GB RAM (GARR) 12 CPU + 64GB RAM (IRE)	12 CPU + 32 GB RAM Workstation	12 TB cold 12 TB hot NAS
2021	12CPU + 64GB RAM (GARR) 32 CPU + 64GB RAM (IRE)	64 CPU 256 GB RAM Workstation	12 TB cold 48 TB hot NAS

Space and time resources





4thNov2020 ISAB

IRE

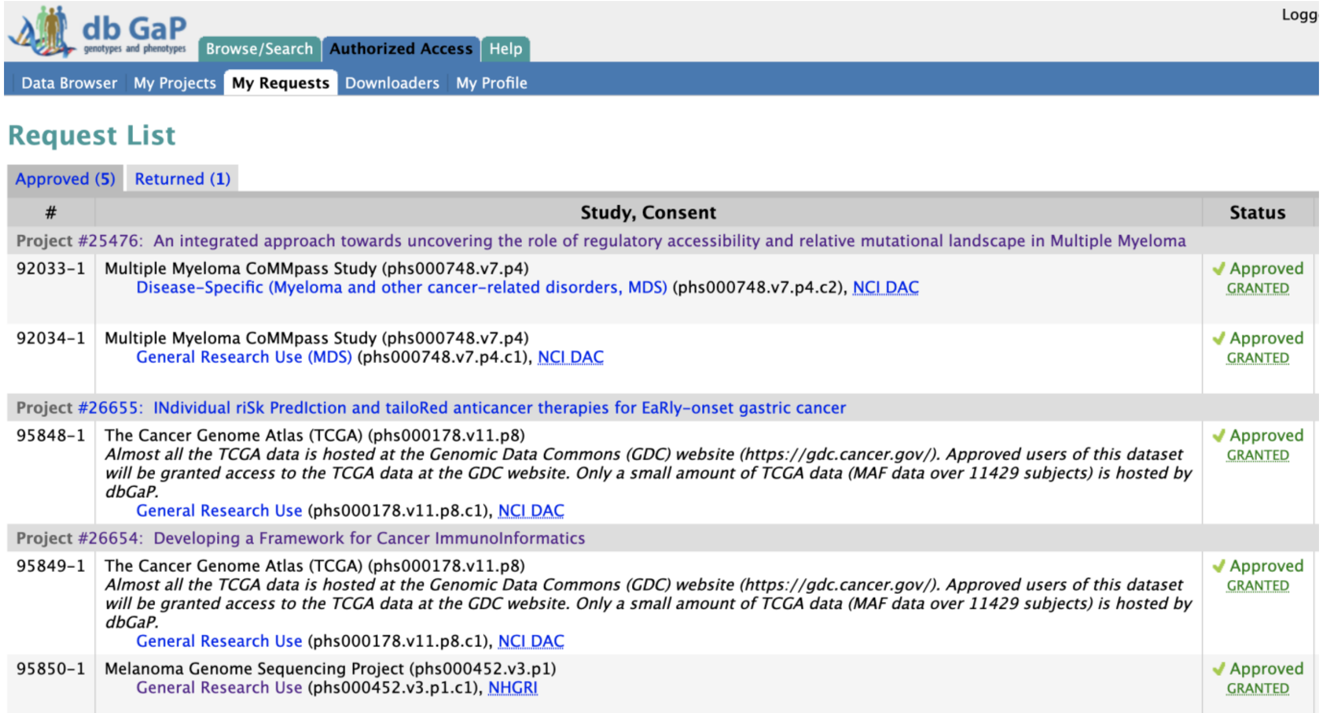
2018: From the International Board report (3)

The Bioinformatics needs to be increased to allow “big” data sets to be properly analyzed and ***include access to major data bases such as TCGA for exploratory studies***

TCGA and Big Data Access

Three main projects/datasets enabled for download and query:

1. *COMPASS Database of 1000 Multiple Myeloma Whole Genome/Exome/RNA profiles*
2. *TCGA Whole Exome Sequencing of Solid Tumors (Panel of Normals to improve tumor-only Variant Calling)*
3. *Melanoma Immune-Checkpoint Treated dataset: WES+RNA-seq*



Request List

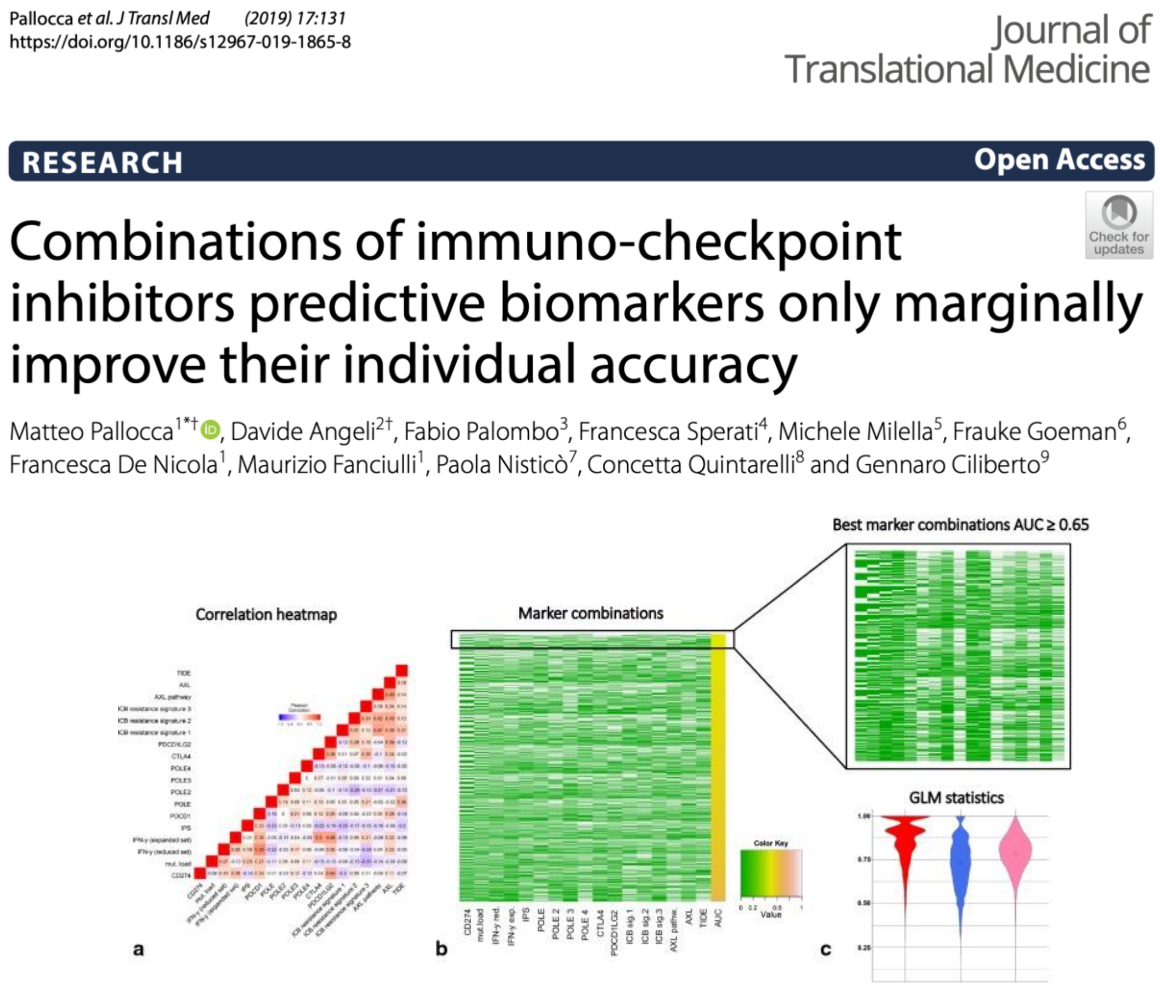
#	Study, Consent	Status
Project #25476: An integrated approach towards uncovering the role of regulatory accessibility and relative mutational landscape in Multiple Myeloma		
92033-1	Multiple Myeloma CoMMpass Study (phs000748.v7.p4) Disease-Specific (Myeloma and other cancer-related disorders, MDS) (phs000748.v7.p4.c2), NCI DAC	✓ Approved GRANTED
92034-1	Multiple Myeloma CoMMpass Study (phs000748.v7.p4) General Research Use (MDS) (phs000748.v7.p4.c1), NCI DAC	✓ Approved GRANTED
Project #26655: INdividual riSk PredIction and tailoRed anticancer therapies for EaRly-onset gastric cancer		
95848-1	The Cancer Genome Atlas (TCGA) (phs000178.v11.p8) *Almost all the TCGA data is hosted at the Genomic Data Commons (GDC) website (https://gdc.cancer.gov/). Approved users of this dataset will be granted access to the TCGA data at the GDC website. Only a small amount of TCGA data (MAF data over 11429 subjects) is hosted by dbGaP.* General Research Use (phs000178.v11.p8.c1), NCI DAC	✓ Approved GRANTED
Project #26654: Developing a Framework for Cancer ImmunoInformatics		
95849-1	The Cancer Genome Atlas (TCGA) (phs000178.v11.p8) *Almost all the TCGA data is hosted at the Genomic Data Commons (GDC) website (https://gdc.cancer.gov/). Approved users of this dataset will be granted access to the TCGA data at the GDC website. Only a small amount of TCGA data (MAF data over 11429 subjects) is hosted by dbGaP.* General Research Use (phs000178.v11.p8.c1), NCI DAC	✓ Approved GRANTED
95850-1	Melanoma Genome Sequencing Project (phs000452.v3.p1) General Research Use (phs000452.v3.p1.c1), NHGRI	✓ Approved GRANTED

Thanks to

Giacomo Corleone
Stefano Scalera
Martina Ferrazzano
Maurizio Fanciulli
Giuseppe Navanteri

Data Science at IRE in 2018-2020: the battle of IPS

Validating and debunking ICI molecular biomarkers



Pallocca *et al. J Transl Med* (2019) 17:131
https://doi.org/10.1186/s12967-019-1865-8

Journal of
Translational Medicine

RESEARCH **Open Access**

Combinations of immuno-checkpoint inhibitors predictive biomarkers only marginally improve their individual accuracy

Matteo Pallocca[1*†], Davide Angeli[2†], Fabio Palombo[3], Francesca Sperati[4], Michele Milella[5], Frauke Goeman[6], Francesca De Nicola[1], Maurizio Fanciulli[1], Paola Nisticò[7], Concetta Quintarelli[8] and Gennaro Ciliberto[9]

4thNov2020

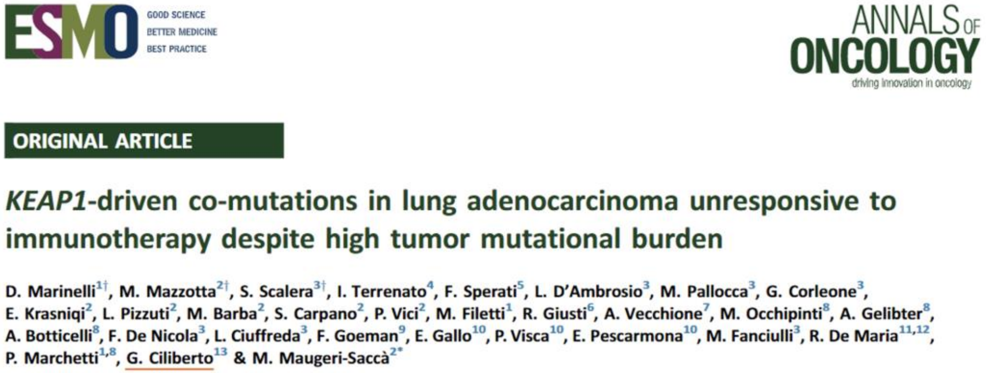
Data Science at IRE in 2018-2020: Oncoinformatics

A group of young Medical Investigators, leveraging the power of:

- Clinical Research
- Statistics + Coding
- Bioinformatics
- Data Modeling
- In-house Molecular data plus Public Datasets

Marcello Maugeri-Saccà,MD

Marco Mazzotta, MD
Stefano Scalera
Daniele Marinelli



ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

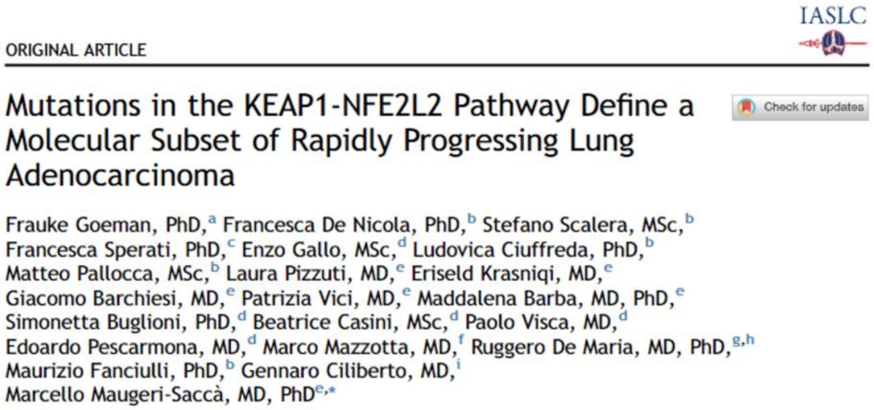
ANNALS OF ONCOLOGY driving innovation in oncology

ORIGINAL ARTICLE

***KEAP1*-driven co-mutations in lung adenocarcinoma unresponsive to immunotherapy despite high tumor mutational burden**

D. Marinelli[1†], M. Mazzotta[2†], S. Scalera[3†], I. Terrenato[4], F. Sperati[5], L. D'Ambrosio[3], M. Pallocca[3], G. Corleone[3], E. Krasniqi[2], L. Pizzuti[2], M. Barba[2], S. Carpano[2], P. Vici[2], M. Filetti[1], R. Giusti[6], A. Vecchione[7], M. Occhipinti[8], A. Gelibter[8], A. Botticelli[8], F. De Nicola[3], L. Ciuffreda[3], F. Goeman[9], E. Gallo[10], P. Visca[10], E. Pescarmona[10], M. Fanciulli[3], R. De Maria[11,12], P. Marchetti[1,8], G. Ciliberto[13] & M. Maugeri-Saccà[2*]

Annals of Oncology 2020
JTO 2020
JITC 2020



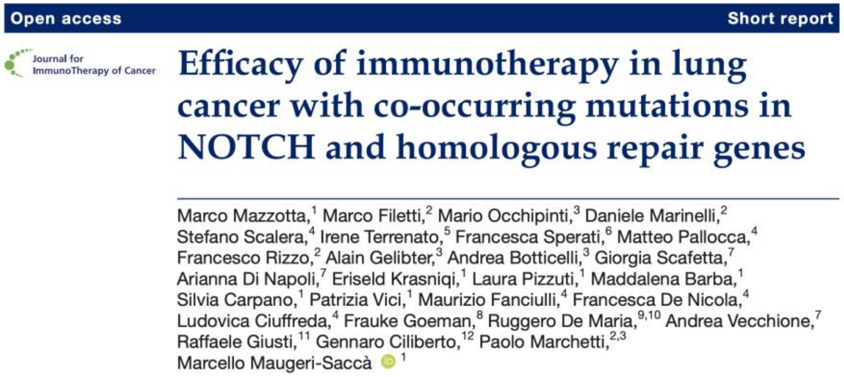
IASLC

ORIGINAL ARTICLE

Check for updates

Mutations in the KEAP1-NFE2L2 Pathway Define a Molecular Subset of Rapidly Progressing Lung Adenocarcinoma

Frauke Goeman, PhD,[a] Francesca De Nicola, PhD,[b] Stefano Scalera, MSc,[b] Francesca Sperati, PhD,[c] Enzo Gallo, MSc,[d] Ludovica Ciuffreda, PhD,[b] Matteo Pallocca, MSc,[b] Laura Pizzuti, MD,[e] Eriseld Krasniqi, MD,[e] Giacomo Barchiesi, MD,[e] Patrizia Vici, MD,[e] Maddalena Barba, MD, PhD,[e] Simonetta Buglioni, PhD,[d] Beatrice Casini, MSc,[d] Paolo Visca, MD,[d] Edoardo Pescarmona, MD,[d] Marco Mazzotta, MD,[f] Ruggero De Maria, MD, PhD,[g,h] Maurizio Fanciulli, PhD,[b] Gennaro Ciliberto, MD,[i] Marcello Maugeri-Saccà, MD, PhD[e,*]



Open access | Short report

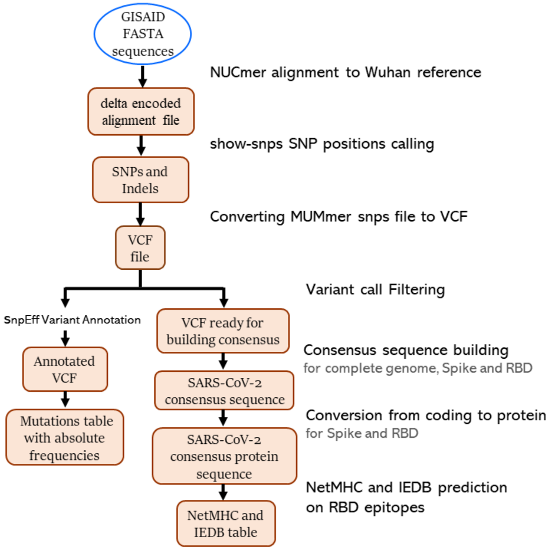
Journal for ImmunoTherapy of Cancer

Efficacy of immunotherapy in lung cancer with co-occurring mutations in NOTCH and homologous repair genes

Marco Mazzotta,[1] Marco Filetti,[2] Mario Occhipinti,[3] Daniele Marinelli,[2] Stefano Scalera,[4] Irene Terrenato,[5] Francesca Sperati,[6] Matteo Pallocca,[4] Francesco Rizzo,[2] Alain Gelibter,[3] Andrea Botticelli,[3] Giorgia Scafetta,[7] Arianna Di Napoli,[7] Eriseld Krasniqi,[1] Laura Pizzuti,[1] Maddalena Barba,[1] Silvia Carpano,[1] Patrizia Vici,[1] Maurizio Fanciulli,[4] Francesca De Nicola,[4] Ludovica Ciuffreda,[4] Frauke Goeman,[8] Ruggero De Maria,[9,10] Andrea Vecchione,[7] Raffaele Giusti,[11] Gennaro Ciliberto,[12] Paolo Marchetti,[2,3] Marcello Maugeri-Saccà[1]

IRE

Data Science at IRE in 2018-2020: Covid-Miner



GISAID FASTA sequences

NUCmer alignment to Wuhan reference

delta encoded alignment file

show-snps SNP positions calling

SNPs and Indels

Converting MUMmer snps file to VCF

VCF file

Variant call Filtering

SnpEff Variant Annotation

Annotated VCF

Mutations table with absolute frequencies

VCF ready for building consensus

Consensus sequence building
for complete genome, Spike and RBD

SARS-CoV-2 consensus sequence

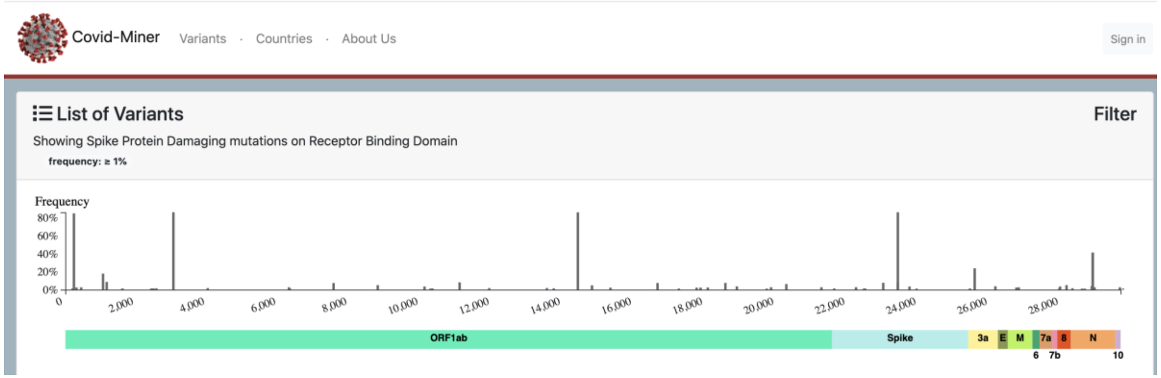
Conversion from coding to protein
for Spike and RBD

SARS-CoV-2 consensus protein sequence

NetMHC and IEDB prediction
on RBD epitopes

NetMHC and IEDB table

https://covid-miner.ifo.gov.it



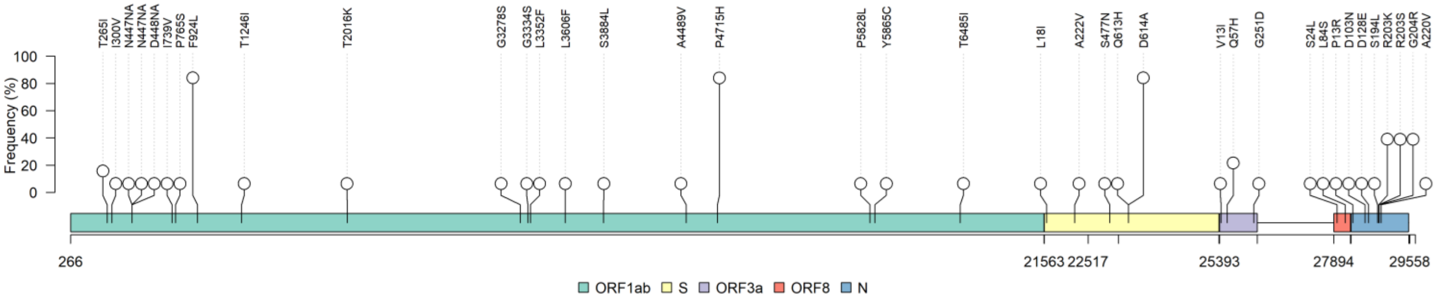
List of Variants

Filter

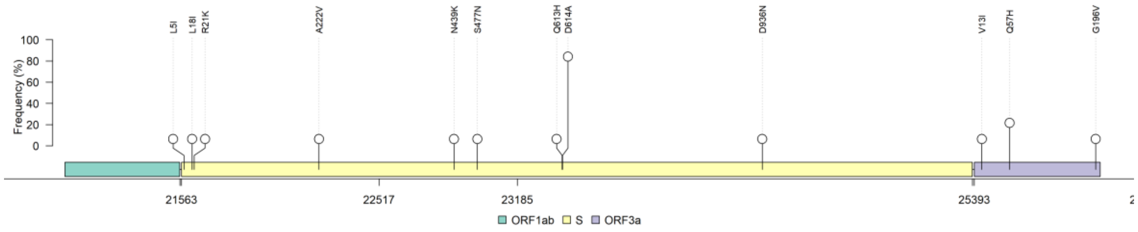
Showing Spike Protein Damaging mutations on Receptor Binding Domain

frequency: ≥ 1%

All Sars-CoV-2 Variation > 1% Frequency

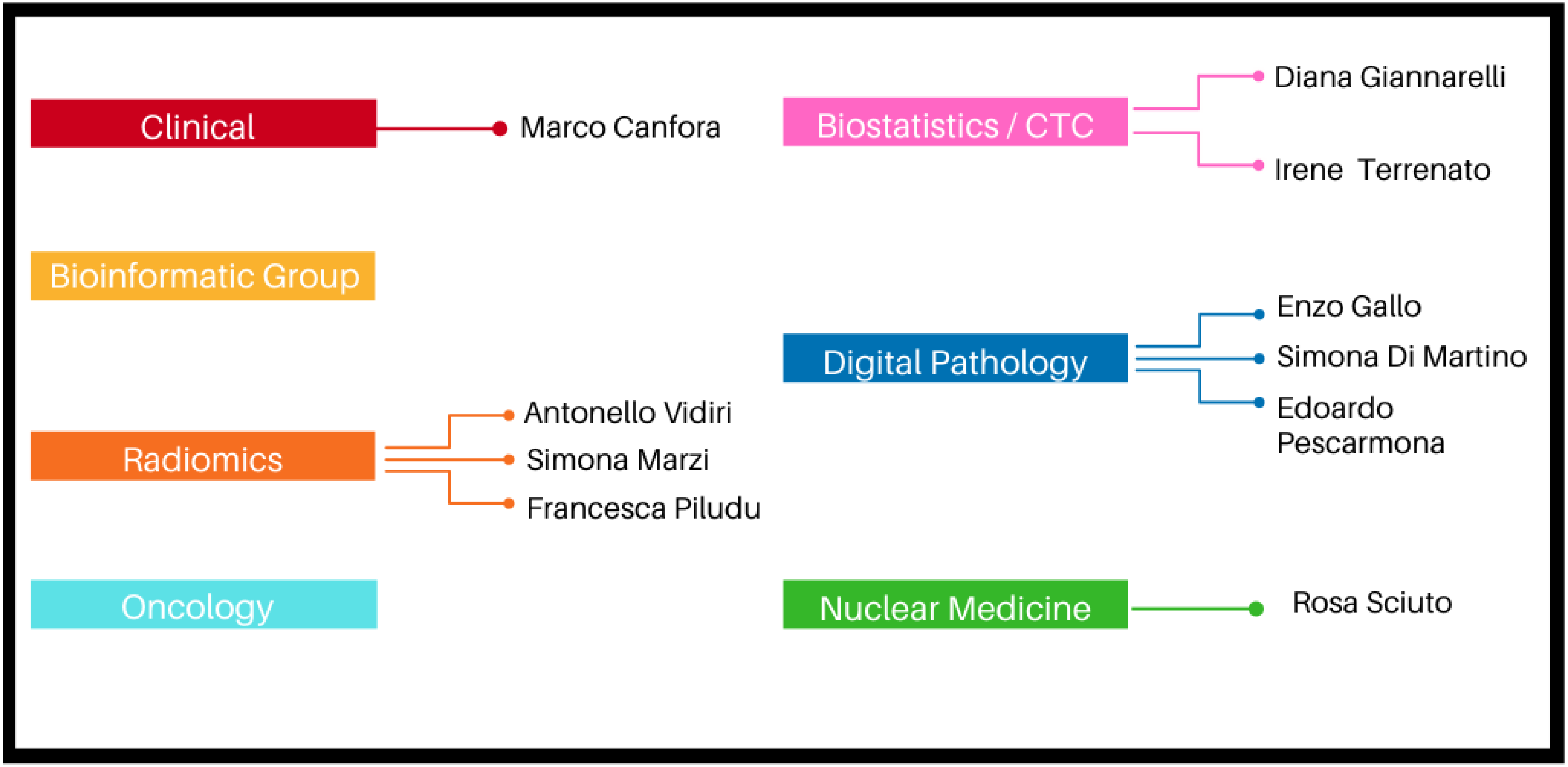


All variations > 0.5% frequency SPIKE



Alice Massacci
Eleonora Sperandio
Lorenzo D’Ambrosio

The Translational Group for **Artificial Intelligence and Imaging**



Clinical
- Marco Canfora

Bioinformatic Group

Radiomics
- Antonello Vidiri
- Simona Marzi
- Francesca Piludu

Oncology

Biostatistics / CTC
- Diana Giannarelli
- Irene Terrenato

Digital Pathology
- Enzo Gallo
- Simona Di Martino
- Edoardo Pescarmona

Nuclear Medicine
- Rosa Sciuto

AI Ongoing: Digital Pathology

Ongoing Digital Pathology Projects

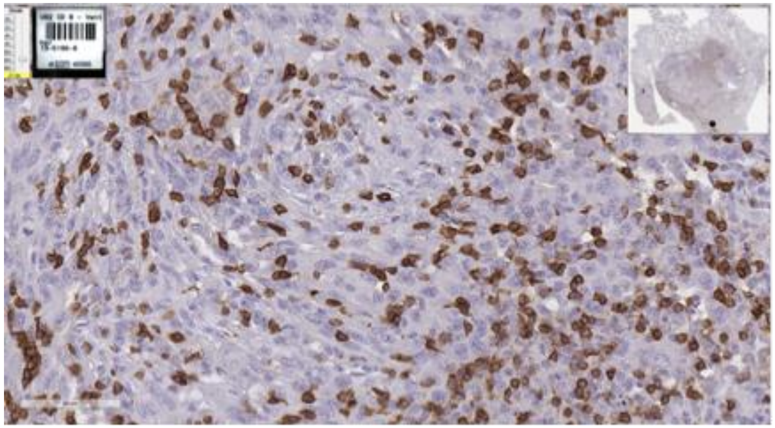
IMMUNOSCORE IN N0 NSCLC PRIMARY TUMORS:
Over 600 immunohistochemistry (IHC) digitized slides under the framework of ACC WG-Immunotherapy.

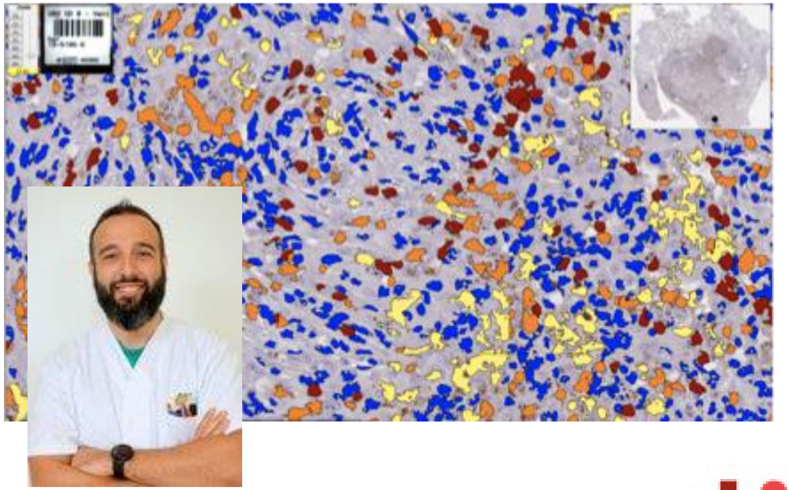
Tumor Infiltrating Lymphocyte in ORL tumor:
50 immunohistochemistry (IHC) digitized slides.

BBIRE/EORTC:
Digitization of histological slides (relating to Biobank samples).

Aperio Genie – System updated for faster image segmentation/ AI models creation.

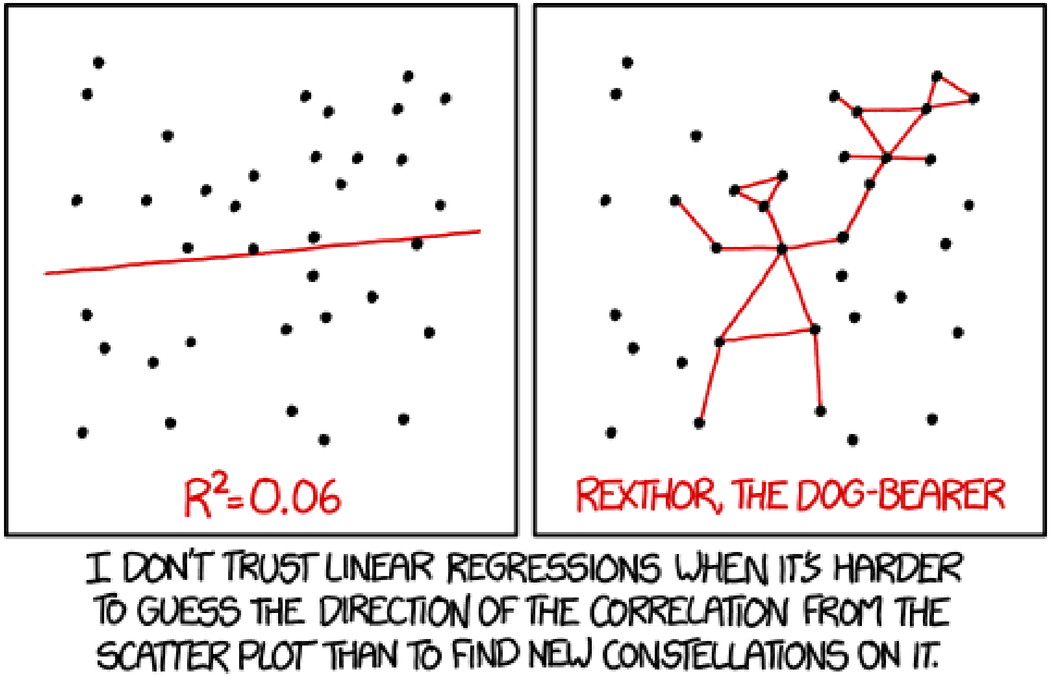
Future: integration with molecular profiles





Enzo Gallo

4thNov2020


R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Thank you for your attention!