RESUM DE RESULTATS OBTINGUTS AL PROJECTE EOSC-LIFE

**Àngel Blanco García i Alex Sánchez Pla**

# Antecedents

En els escenaris de recerca Biomèdica actuals, aspectes tan importants com el coneixement adequat de la provinença, traçabilitat i transformació de dades, els mètodes emprats i l’exactitud en els paràmetres de les anàlisis, i la presentació acurada de resultats, es poden resentir greument per mancances d’interoperabilitat en els fluxos de treball computacionals. Addicionalment, entorns relativament nous com la computació al núvol, amb els seus clars avantatges, poden contribuir encara més a aquests problemes si impliquen eines i mètodes propietaris o poc oberts. Donar molt més protagonisme a la interoperabilitat i la transparència en el funcionament són les vies principals per mitigar aquests problemes[1], [2], [3], [4]. Aquesta tasca no és en absolut trivial ni fàcilment assolible.

Per facilitar aquesta transició en l'àmbit de les Biociències, cal un enfocament amb vocació generalista i sistemàtic de com implementar recursos computacionals de manera interoperable, basat en estandars oberts i amb especial atenció als desplegaments al núvol pel seu potencial multiplicador de la interoperabilitat, que pugui facilitar la conversió d’eines computacionals existents i guiar-ne la creació de noves. El projecte **EOSC-Life**[5] es va posar en marxa per crear i fer disponibles els mitjans per assolir aquest objectiu a nivell paneuropeu. És projecte de gran abast que cobreix des de la descripció i ús de dades i serveis (seguint els principis **FAIR**[6]) fins a a la infraestructura per desplegament i el *training*, passant per la creació o re-implementació de *workflows* i el tractament de dades sensibles a nivell de privacitat.

En concret, el ***Work Package* 2** del projecte[7] té com a principal objectiu possibilitar la creació o adaptació d’eines de software per tal que siguin interoperables en entorns de núvol de computació oberts, mitjançant la implementació basada en un conjunt estandaritzat d’eines i mètodes bàsics[8] que fa aquesta interoperabilitat possible (anomenat internament ***WP2 Roadmap***). Actualment participo en aquest *Work Package* mitjançant la meva feina com a membre de la **Unitat d’Estadística i Bioinformàtica** (UEB) del Vall d’Hebron Institut de Recerca (VHIR).

Com a part d’aquest treball, s’ha escollit el *pipeline* **DOMINO**[9] per dur a terme la seva conversió a *workflow* interoperable per entorns *open cloud*, de manera alineada amb l’esmentada participació en el *Work Package* 2 del projecte *EOSC-Life*. DOMINO va ser desenvolupat en la seva forma original al **Evolutionary Genomics and Bioinformatics Group** del Departament de Genètica, Microbiologia i Estadística de la Universitat de Barcelona. Es tracta d’un grup capdavanter en el seu camp[10], [11], [12], [13], i aquest vincle ofereix la possibilitat tant de fer la re-implementació esmentada de DOMINO, com de poder aprofitar l’experiència i coneixements del grup per utilitzar l’eina per anàlisis i estudis rellevants una vegada finalitzada la seva re-implementació.

# Objectius

L’objectiu general de la nostra participació ha estat crear un entorn de computació en núvol obert (*open- cloud*) on poder desplegar *workflows* de tipus *open-cloud ready* implementats seguint les directrius establertes pel *Work Package* 2 del projecte EOSC-Life. El cas d'ús principal ha estat el desenvolupament a partir d’un pipeline existent, DOMINO, i el procés inclou totes les fases, des de la preparació del sistema com a *workflow*, el desplegament al núvol i, finalment, la seva utilització per portar a terme análisis de dades rellevants (amb dades de seqüenciació massiva) provinents de grups clínics o de recerca d’àmbits propers o relacionats amb els descrits a l’apartat d’Antecedents (Vall d’Hebron / VHIR i Evolutionary Genomics and Bioinformatics Group de la UB).

Per assolir aquest objectiu general, es van establir tres objectius principals, que es citen a continuació:

1. **Re-implementació del pipeline DOMINO com a *workflow* de tipus *open-cloud ready***
2. **Creació d’un entorn de computació en núvol obert**
3. **Anàlisi de dades**

En el període cobert per aquest resum s’ha treballat en tasques corresponents **als dos primers objectius**, a. Més en concret, per a cada un dels dos objectius, s’ha treballat en:

Objectiu principal 1:

* Re-factorització dels mòduls de *software* que intregren l’eina
* Afegir nova funcionalitat a DOMINO
* *Containerització* dels blocs rellevants del pipeline
* Implementació per a *Workflow Management Systems*

Objectiu principal 2:

* Creació i posada en marxa d’una instància de **Galaxy**[14] local (àmbit Vall d’Hebron Institut de Recerca)

El treball dut a terme durant aquest període pels dos objectius es desenvolupa més en detall en els apartats de Metodologia i Resultats.

# Metodologia

**Dades:**

Per avaluar els canvis que s’han anat introduint a DOMINO, s’han utilitzat com a conjunts de prova dades provinents de diferents mètodes de seqüenciació (454, *Illumina*, *PacBio*) disponibles al Evolutionary Genomics and Bioinformatics Group del Departament de Genètica, Microbiologia i Estadística de la Universitat de Barcelona.

**Eines computacionals:**

Donat que aquest projecte implica una considerable integració de diferents àmbits computacionals, el conjunt d’eines utilitzat és força ampli. A continuació esmentem algunes de les més representatives, especialment pel que fa a tasques d'anàlisi i desenvolupament de *software* i d’operacions de sistemes, tant virtualitzats com físics.

* Desenvolupament: Python, Perl, Groovy, C++, Nextflow DSL (versions 1 i 2), i mòduls i eines de desenvolupament associades.
* Contenidors: Singularity, Docker, Podman, LXD i varies de les seves eines associades de construcció, desplegament i orquestració.
* Configuració, Provisionament, Orquestració: Ansible, WareWulf, Vagrant, Terraform
* Marc integrat per sistemes HPC: OpenHPC
* Gestor de tasques computacionals (HPC): SLURM

S’han utilitzat moltes altres eines auxiliars o amb papers més puntuals, en base a les necessitats existents de cada etapa de treball.

# Resultats

### Objectiu principal 1: Re-implementació del pipeline DOMINO com a workflow de tipus *open-cloud ready*

* **Re-factorització dels mòduls de *software* que intregren l’eina**: S’ha analitzat la implementació existent de DOMINO per identificar i posteriorment millorar etapes en funció de criteris com rendiment, estabilitat, modularitat i interoperabilitat, i sostenibilitat del *workflow*. Les millores han implicat diferents accions en funció de la situació i natura de cada una de les etapes de DOMINO en les que s’ha treballat. En ocasions s’ha intervingut re-implementant etapes existents. En altres ocasions s’han substituït aquestes etapes per altres anàlogues incorporant eines o conjunts d’eines diferents. Aquest procés no està finalitzat, ja que hi ha etapes de DOMINO per les quals encara es considera que hi ha marge de millora raonable, i amb algunes de les opcions possibles en avaluació i discussió.
* **Afegir nova funcionalitat a DOMINO:** S’ha treballat en la integració de l’ensamblador *de novo* per *short reads* **ABySS**[15], i de mòduls per dotar a DOMINO de la capacitat de treballar amb ensamblatges de transcriptomes provinents de *RNA-Seq*. Per ambdues tasques la metodologia s’ha basat d’entrada en proveir aquestes noves funcionalitats (i d’altres en el futur) de manera modular amb alt grau d’interoperabilitat i flexibilitat integració amb els altres elements del *workflow*. Com a resultat d’aquest procés, l’ensamblador **ABySS** està actualment integrat en DOMINO. Es continua treballant en la integració de mòduls per utilitzar ensamblatges de transcriptomes provinents de *RNA-Seq*. Amb el treball que s’ha fet fins al moment, la majoria d’etapes necessàries per assolir aquesta tasca estan o bé finalitzades i integrades en DOMINO o en estats propers.
* ***Containerització* dels blocs rellevants del pipeline**: S’ha fet un anàlisi dels contenidors requerits i de la seva disposició en el *workflow*. Aquest contenidors s’han construït en els casos necessaris, o bé s’han utilitzat contenidors ja disponibles amb les eines requerides[16],[17], sempre que estiguin dissenyats i construïts de manera que es puguin integrar adequadament en un workflow *open-cloud ready* com el que és el nostre objectiu. Totes les etapes de DOMINO incloses fins al moment ja estan implementades per treballar amb contenidors de la manera indicada.
* **Implementació per a *Workflow Management Systems:*** Inicialment s’ha optat per començar a implementar els canvis, millores, i modularitzacions detallats en els tres punts anteriors en un *workflow* de **Nextflow**[18]**.** Aquest *workflow management system* permet al desenvolupador una implementació més àgil de prototips del *workflow* (totals o parcials), i per tant és més adient per les etapes inicials de desenvolupament del projecte.

### Objectiu principal 2: Creació d’un entorn de computació en núvol obert

S’ha treballat en la creació d’una instància de Galaxy utilitzant infraestructura de la Unitat d’Estadística i Bioinformàtica del Vall d’Hebron Institut de Recerca (VHIR). Aquest treball implica importants components de coneixement i ús de metodologies d’administració i configuració de sistemes, així com de tasques integrades de desenvolupament i operacions (*DevOps)*. A més, donat que ens proposem que aquesta instància de Galaxy pugui donar servei al VHIR quan assoleixi l’estat de producció, és necessari que serveixi d’interface amb una infraestructura de computació dedicada, com per exemple un clúster. És per això que aquesta tasca també ha requerit de treball d'administració i configuració de sistemes HPC, ja que en el nostre cas s’utilitzarà el clúster de la Unitat d'Estadística i Bioinformàtica del VHIR.

# Conclusions

* S’ha dut a terme una refactorització extensa dels mòduls existents de DOMINO com a part essencial per convertir-lo en un *workflow* *open cloud ready,* procés que continua per certs mòduls rellevants en els que s’aprecia marge per millora i/o innovació.
* S’han afegit gran part de les noves funcionalitats de DOMINO que s’havien projectat per aquest període.
* DOMINO actualment ja utilitza eines *containeritzades* per totes les etapes rellevants. Les possibles millores i modificacions de DOMINO parteixen d’utilitzar també aquesta *containeritzacio* com a premisa. Aquest és un pas essencial per assolir que DOMINO sigui un *workflow open cloud ready*.
* La implementació de DOMINO en un *Workflow Management System* s’ha materialitzat inicialment amb Nextflow, ja que afavoreix la creació i avaluació àgil i ràpida de prototips. Gran part de les modificacions que s'implementen actualment en DOMINO es fan de manera directa en Nextflow.
* S’ha creat uns instància de Galaxy en la infraestructura del VHIR, que actualment està en estat de pre-producció.

# Referències

1. S. Möller, S.W. Prescott, L. Wirzenius, P. Reinholdtsen, B. Chapman, P. Prins, et al., Robust Cross-Platform Workflows: How Technical and Scientific Communities Collaborate to Develop, Test and Share Best Practices for Data Analysis, *Data Sci. Eng.* 2 (2017) 232–244.
2. B. Grüning, J. Chilton, J. Köster, R. Dale, N. Soranzo, M. van den Beek, et al., Practical computational reproducibility in the life sciences., *Cell Syst.* 6 (2018) 631–635.
3. D. Garijo, Y. Gil, O. Corcho, Abstract, link, publish, exploit: An end to end framework for workflow sharing, *Future Generation Computer Systems*. 75 (2017) 271–283.
4. P.D. Schloss, Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research., *MBio.* 9 (2018).
5. <https://cordis.europa.eu/project/id/824087>
6. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; et al. (15 March 2016). "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*. **3**: 160018.
7. <https://forum.eosc-life.eu/t/introduction-to-eosc-life-wp2/32>
8. <https://forum.eosc-life.eu/t/eosc-life-wp2-roadmap/31>
9. Cristina Frías-López, José F. Sánchez-Herrero, Sara Guirao-Rico, Elisa Mora, Miquel A. Arnedo, Alejandro Sánchez-Gracia, Julio Rozas. DOMINO: development of informative molecular markers for phylogenetic and genome-wide population genetic studies in non-model organisms, *Bioinformatics*, Volume 32, Issue 24, 15 December 2016, Pages 3753–3759
10. Jose Francisco Sánchez-Herrero, Cristina Frías-López, Paula Escuer, Silvia Hinojosa-Alvarez, Miquel A Arnedo, Alejandro Sánchez-Gracia, Julio Rozas, The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates, *GigaScience*, Volume 8, Issue 8, August 2019, giz099
11. Joel Vizueta, Paula Escuer, Cristina Frías-López, Sara Guirao-Rico, Lars Hering, Georg Mayer, Julio Rozas, Alejandro Sánchez-Gracia, Evolutionary History of Major Chemosensory Gene Families across Panarthropoda, *Molecular Biology and Evolution*, Volume 37, Issue 12, December 2020, Pages 3601–3615
12. Vizueta, J., Sánchez‐Gracia, A., & Rozas, J. (2020). BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *Molecular Ecology Resources*, 20, 1445– 1452
13. Almudi, I., Vizueta, J., Wyatt, C.D.R. *et al.* Genomic adaptations to aquatic and aerial life in mayflies and the origin of insect wings. *Nat Commun* 11, 2631 (2020).
14. Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, *Nucleic Acids Research*, Volume 46, Issue W1, 2 July 2018, Pages W537–W544, doi:10.1093/nar/gky379
15. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, 2017 27: 768-777
16. da Veiga Leprevost F, Grüning BA, Alves Aflitos S, Röst HL, Uszkoreit J, Barsnes H, Vaudel M, Moreno P, Gatto L, Weber J, Bai M, Jimenez RC, Sachsenberg T, Pfeuffer J, Vera Alvarez R, Griss J, Nesvizhskii AI, Perez-Riverol Y. BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics*. 2017 Aug 15;33(16):2580-2582. doi: 10.1093/bioinformatics/btx192.
17. <https://depot.galaxyproject.org/singularity/>
18. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), 316–319.
19. Kristoffer Sahlin. Effective sequence similarity detection with strobemers. Genome Res. 2021. 31: 2080-2094
20. Hua, Z., Jiang, C., Song, S., Tian, D., Chen, Z., Jin, Y., Zhao, Y., Zhou, J., Zhang, Z., Huang, L., & Yuan, Y. (2022). Accurate identification of taxon-specific molecular markers in plants based on DNA signature sequence. Molecular Ecology Resources, 00, 1– 12. https://doi.org/10.1111/1755-0998.13697
21. Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso & Sven Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology volume 38, 276–278 (2020)