GREEKC

GREEKC

Gene Regulation Ensemble Effort for the Knowledge Commons

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

NTNU – Trondheim
Norwegian University of
Science and Technology

COST Action CA15205

GREEKC:

Gene Regulation Ensemble Effort for the Knowledge Commons

a.k.a.

GRECO:

Gene Regulation Consortium

Action Chair: Martin Kuiper

www.greekc.org

GREEKC

GREEKC

Gene Regulation Ensemble Effort for the Knowledge Commons

The Knowledge Commons

the Knowledge Commons:

“the collection of freely accessible information resources, with data well annotated with unambiguous descriptors according to quality criteria and standards that allow seamless integration and interoperability as well as automated computational access with third party software”

Main aim

Provide help to structure the ‘Gene Regulation Knowledge Commons’

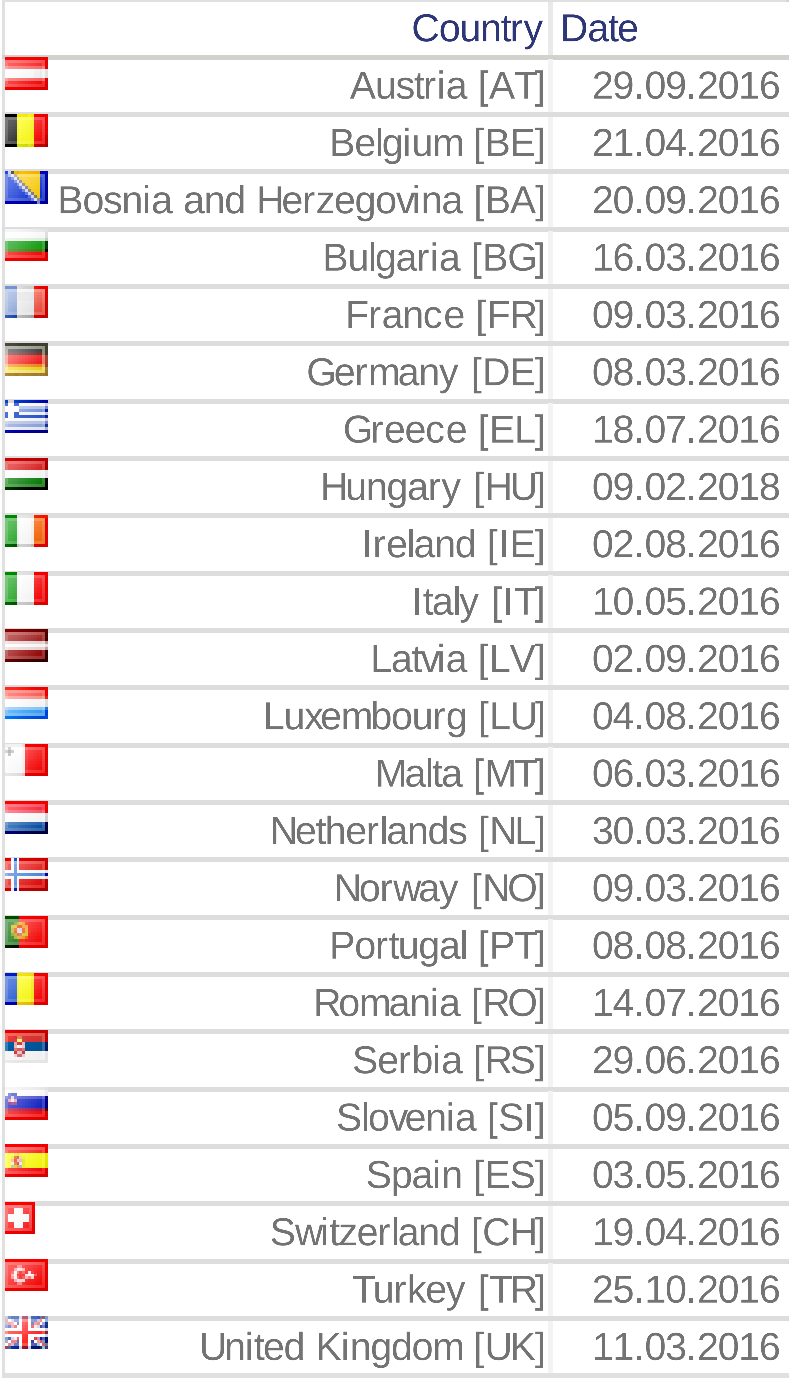
What is the Gene Regulation Knowledge Commons (GRKC)?

“the set of web databases/knowledge bases that contain everything we know about gene regulation processes”

Our consortium is organised in four Working Groups (WGs), focusing on Ontology development, Curation approaches, Text mining, and Databases.

We organise Training Events and Workshops to review the state-of-the-art, discuss hurdles and bottlenecks, and define the way forward

The consortium:

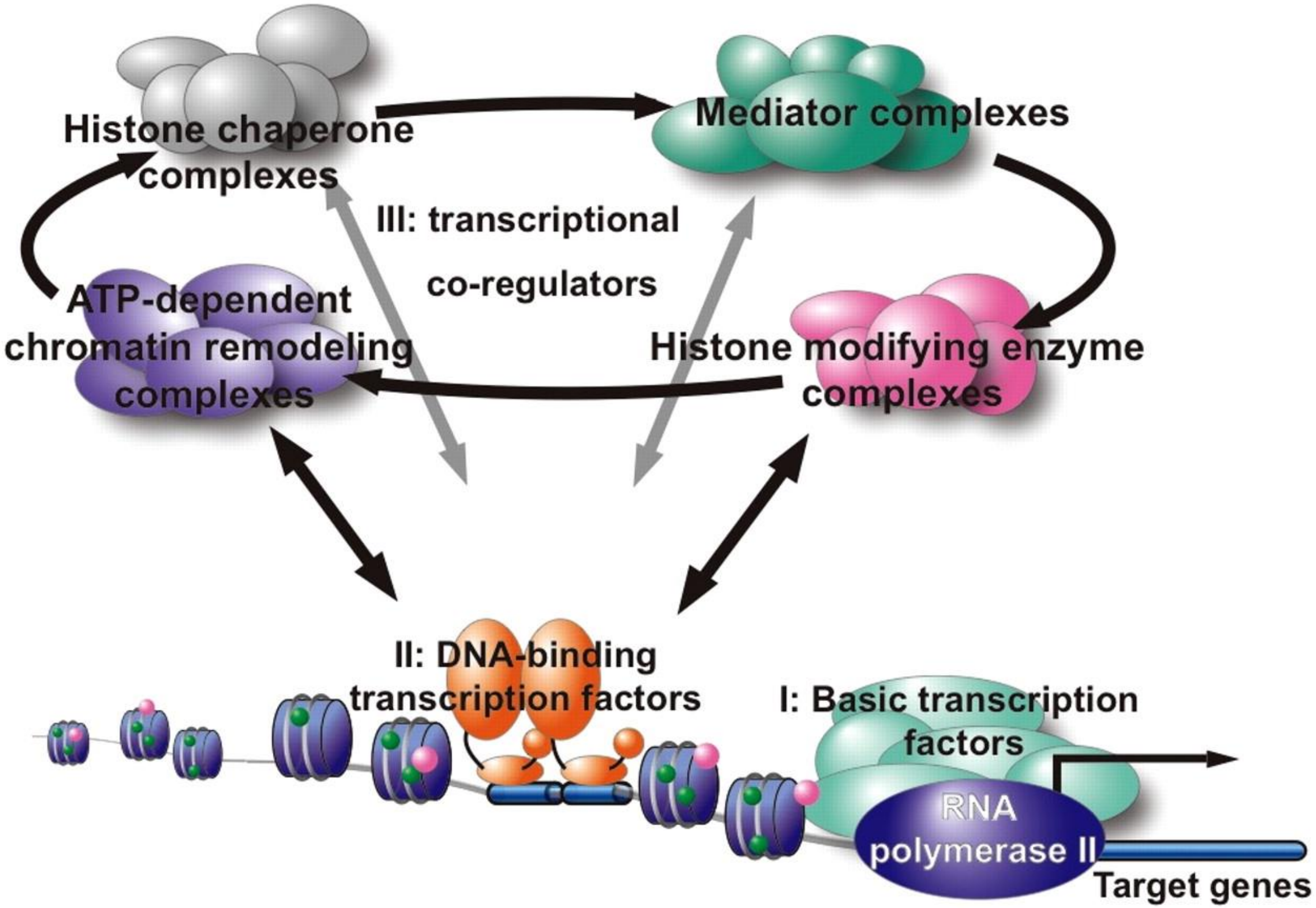


Country	Date
Austria [AT]	29.09.2016
Belgium [BE]	21.04.2016
Bosnia and Herzegovina [BA]	20.09.2016
Bulgaria [BG]	16.03.2016
France [FR]	09.03.2016
Germany [DE]	08.03.2016
Greece [EL]	18.07.2016
Hungary [HU]	09.02.2018
Ireland [IE]	02.08.2016
Italy [IT]	10.05.2016
Latvia [LV]	02.09.2016
Luxembourg [LU]	04.08.2016
Malta [MT]	06.03.2016
Netherlands [NL]	30.03.2016
Norway [NO]	09.03.2016
Portugal [PT]	08.08.2016
Romania [RO]	14.07.2016
Serbia [RS]	29.06.2016
Slovenia [SI]	05.09.2016
Spain [ES]	03.05.2016
Switzerland [CH]	19.04.2016
Turkey [TR]	25.10.2016
United Kingdom [UK]	11.03.2016

Associated countries:

Russia
Ukraine
USA
Canada
Mexico
Brazil
Japan

Gene expression regulation is a complex biological domain



Histone chaperone complexes
Mediator complexes
III: transcriptional co-regulators
ATP-dependent chromatin remodeling complexes
Histone modifying enzyme complexes
II: DNA-binding transcription factors
I: Basic transcription factors
RNA polymerase II
Target genes

From
Emergence of the osteo-epigenome in bone biology
Shigeaki Kato, Kazuki Inoue & Min-Young Youn
IBMS BoneKEy (2010) **7**, 314–324 (2010) | doi:10.1138/20100464
Published online September 2010

What would be needed to standardise the field ?

- Curation:
  - The annotation of biological entities and relationships that are core to regulation of gene expression, at protein, RNA, DNA, 2D and 3D level
- Ontologies to support annotation and subsequent querying of the knowledge commons
  - GO, SO, RO, PSI-MI, GRO, GRAO…
- Curation guidelines to guide annotation
  - Quality and proof levels
  - Provenance
- Harvesting text to streamline the curation process
  - Triage, classifiers, pre-markings of entities and relationships
- Databases and exchange mechanisms
  - IntAct, Signor, Uniprot, GOA; PSI-MI, CausalTAB, PSICQUIC, …

GREEKC/GRECO Objectives

- **Main Objectives:**
- Structure the archiving of information about regulation of gene expression
- Develop guidelines for recording information
- Build and extend ontologies for the annotation of molecules, complexes, interactions and motifs
- Engage the community in the curation of knowledge in literature, introduce text mining in curation workflow
- Upload knowledge to primary and secondary databases

Working Groups

WG1: Development of ontologies and controlled vocabularies

WG2: Development of curation guidelines and standards for five subdomains:

a) Gene Regulators at the protein level (e.g. transcription factors (TF))

b) Gene Regulators at the RNA level (e.g. ncRNAs)

c) The nucleotide sequence recognition level (TF binding sites)

d) The genome level (e.g. methylation status, histone modifications, etc.)

e) The interaction level (e.g. TF complexes with DNA)

WG3: Text mining for automated finding of gene regulation information

WG4: Databasing - storing and sharing of annotations

How do we organise our work?

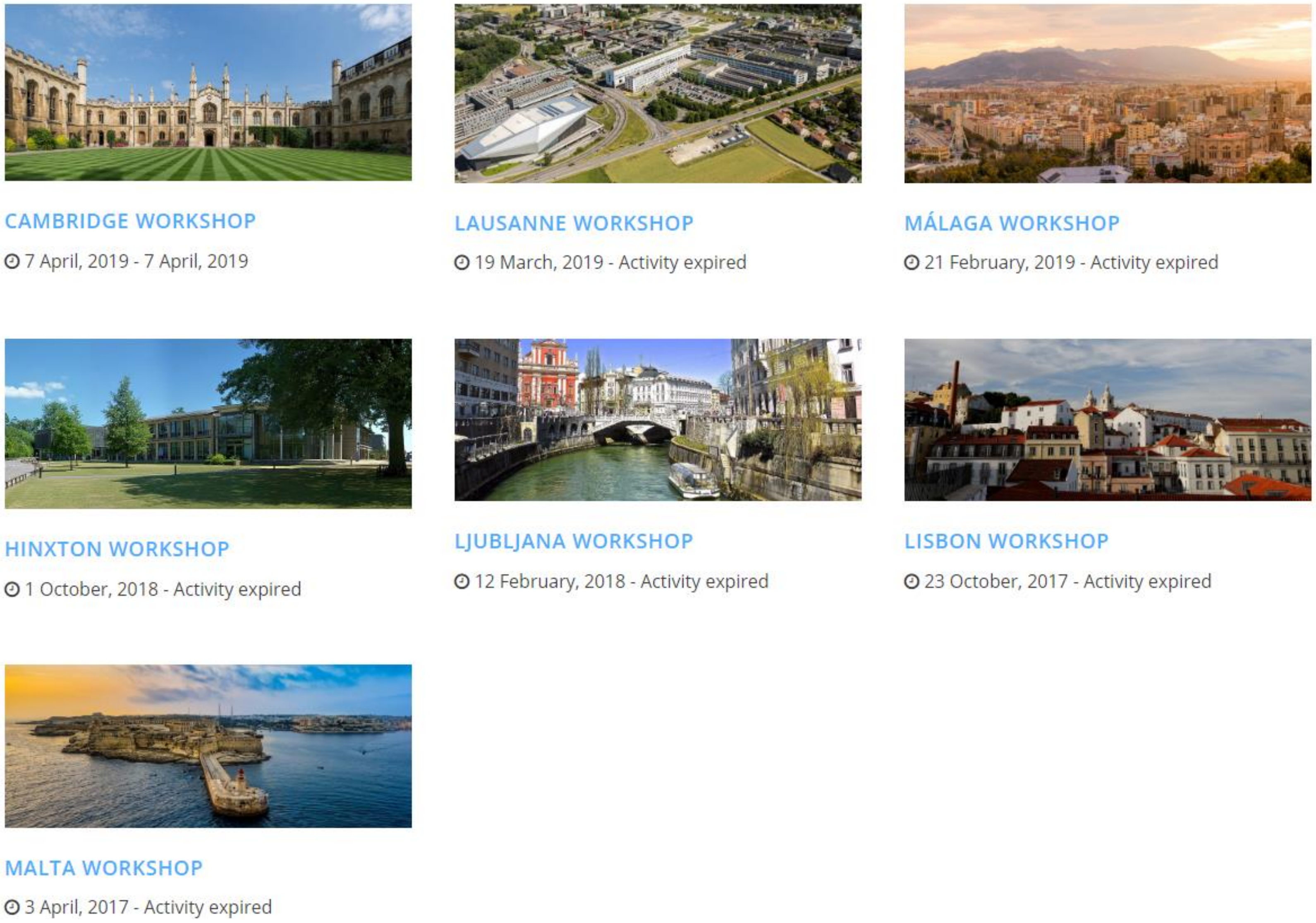
We have different ‘instruments’ available:

MC meetings, Workshops, Training Schools, STSMs

- MC meetings: administrative, touch base with management committee (some 50 members)
- Workshops: this is where the ‘action’ is: discussion about Working Group issues
- Training Schools: essentially a test of where we are with the interoperable knowledge commons
- STSMs: short term scientific missions, for early career scientists to work on GREEKC issues

How do we organise our work (2)?

Workshops:



CAMBRIDGE WORKSHOP

7 April, 2019 - 7 April, 2019

LAUSANNE WORKSHOP

19 March, 2019 - Activity expired

MÁLAGA WORKSHOP

21 February, 2019 - Activity expired

HINXTON WORKSHOP

1 October, 2018 - Activity expired

LJUBLJANA WORKSHOP

12 February, 2018 - Activity expired

LISBON WORKSHOP

23 October, 2017 - Activity expired

MALTA WORKSHOP

3 April, 2017 - Activity expired

How do we organise our work (2)?

Workshops:

WS #1: Malta – Setting the stage for the Gene Regulation Knowledge Commons

- **WG1, WG2, WG3, WG4**

WS #2: Lisbon – Interoperability and the Knowledge Commons

- **WG1, WG2, WG3, WG4**

WS #3: Ljubljana – Representing and Assessing Transcription Factor Binding Specificity

- **WG1, WG2**

WS #4: Hinxton – Ontologies and FAIR tooling

- **WG1**

WS #5: Málaga – Text mining in curation workflows

- **WG3**

WS #6: Lausanne – DNA centric annotation and Transcription Factor binding motifs

- **WG1, WG2c, d**

WS #7: Cambridge – Stakeholder input

- **WG1, WG2, WG3, WG4**

How do we organise our work (3)?

Training Schools

They have taken the shape of a hackathon: the development of bioinformatic workflows that solve biological questions, test the interoperability of resources and challenge the working groups for future improvements

MARSEILLE TRAINING EVENT

23 April, 2019 - 26 April, 2019

LISBON TRAINING EVENT

25 October, 2017 - Activity expired

How do we organise our work (3)?

Training School Marseille: 23 – 26 April, Aix Marseille University / Luminy

- Agenda: The GREEKC COST Action organises a 3-days training session / hackathon in Marseille, that will aim at developing and using interoperable workflows invoking remote bioinformatics resources (knowledge bases, databases, analysis tools) to gather information about regulatory networks and enable an integrative analysis of gene regulation processes.

Bioinformatics resources relevant for GREEKC:

- Specialized resources for gene regulation (Ensembl Regulation, JASPAR, Regulatory Sequence Analysis Tools, ReMap, ArrayExpress, …);
- Core data resources about gene and protein functions (EnsemblGenomes, Uniprot, …);
- ELIXIR interoperability resources and catalogues (bio.tools, BioSchemas, …);
- Literature knowledge (Europe PMC / PubMed); extracted by text mining: ExTRI corpus - Transcription Factor / Target Gene information generated by NTNU/BSC. (www.BioGateway.eu); IMEX resources / IntAct, Signor):
- Any other relevant resource that would answer specific needs emerging from the hacking sessions.

How do we organise our work (4)?

STSMs:

The GREEKC COST Action provides the possibility for Short-Term Scientific Missions (STSM) fellowships. This support comes in the form of a travel grant for an exchange visit between researchers involved in a COST Action. In particular young scientists from a COST country can visit an institution or laboratory in another COST country. These visits are aimed at fostering collaboration, sharing new techniques and infrastructure that may not be available in other participants' institutions or laboratories. The STSMs in GREEKC are intended especially to foster collaboration in the area of biocuration, data standard and ontology development, and data sharing focusing on the topic of Gene Regulation.

The procedure to apply for it is to contact a Host (see the GREEKC webpage) and together draft a proposal for a small project, which is then submitted here:

https://e-services.cost.eu/user/login/STSM

Applications are evaluated and prioritised for funding by the STSM coordinator. Typically the STSM is for 2 – 12 weeks, with a grant total between 1000 and 2500 Euro.

Results:

Use cases:

UC1: TF-TG matrixes for inferring TF activity from transcriptor (Astrid Lagreid/Luz Garcia Alonso))

UC2: small RNAs;TF and miRNA co-regulation of TGs (Vesselin Baev)

UC3: cis-regulatory modules (CRMs), genomic coordinates, structure-function (Ivan Kulakovskiy)

UC4: Inference of Gene Regulatory Networks (Maria del Mar)

UC5: Querying the GEXKB (GRAO-type knowledge base) with composite term sets to retrieve subsets of data, imagine updated GO/SO and predict enhancements / differences (Martin Kuiper)

UC6: Epigenetic regions (Colin, Yulia Medvedeva)

UC7: lncRNAs in regulation (Yulia Medvedeva)

UC8: additional use case support on non-coding RNAs (Ivo Grosse)

UC9: multi-omics integration to infer potential TF-TG interactions (Rafel Riudavets Puig)

Results:

Use case 1:

UC1: TFTG matrixes for inferring TF activity from transcriptome (Astrid Lagreid/Luz Garcia Alonso))

Data, knowledge and tools needed:

- Transcriptomes
- TF set, means to assign their binfing to cisregulatory regions
- TF – TG matrix
- Activities of these TFs and TGs in the tissues/cells that provided the transcriptome datasets
- Mathematical framework / tools to determine TF activities that maximally explain the transcriptome

→ To be further developed in Marseille

Results:

UC0.1: What TFs regulate this gene? (Lausanne discussion / Daniel Zerbino)
UC0.2: What genes does this TF regulate? (Lausanne discussion / Daniel Zerbino)

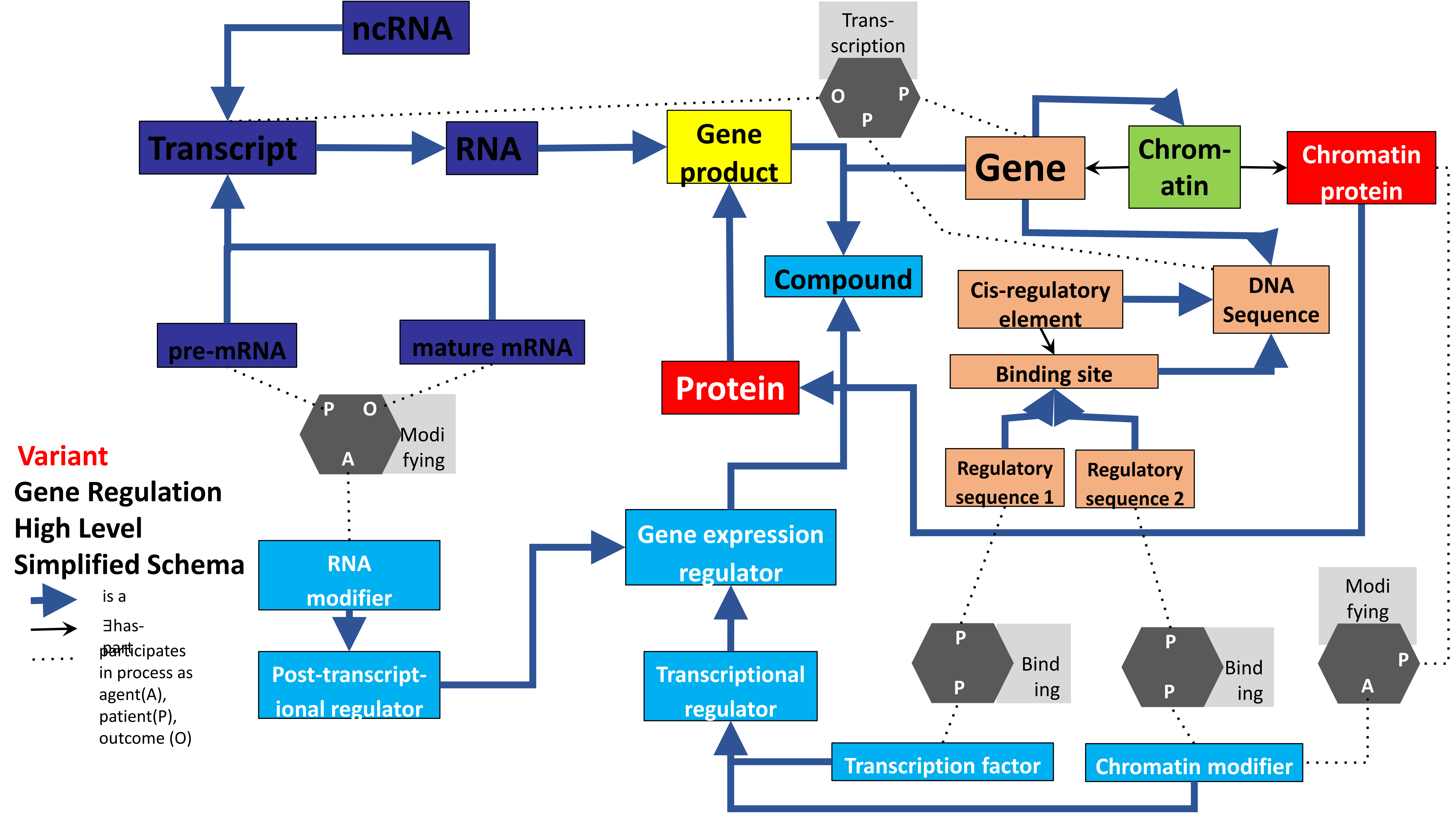
Available data sets for Use case worflows (Lausanne workshop):

- KO experiments - Achilles
- CROP-seq (C. Bock)
- Allele specific binding - GM12878, AlleleDB
- CisMapper - correlation based assessment of TF activity
- Loops (ENCODE new data Chia-PET - to be checked for availability (Zaugg))
- TADs (Blueprint)
- Genetics (GTEx eQTL)
- GeneHancer
- Pharma perturbations (UCSC?), CMAP (Connectivity Map) - BROAD
- Expression correlation (Expression Atlas)
- hQTL (Zaugg data)
- Validate by predicting gene expression
- Protein phosphorylation data (e.g. Beltrao)
- Text mine free text from UniProt
- HOCOMOCO, JASPAR, CISBP motif sources?
- ENCODE? - FactorBook
- REMAP
- Resource from Mathelier: https://unibind.uio.no/

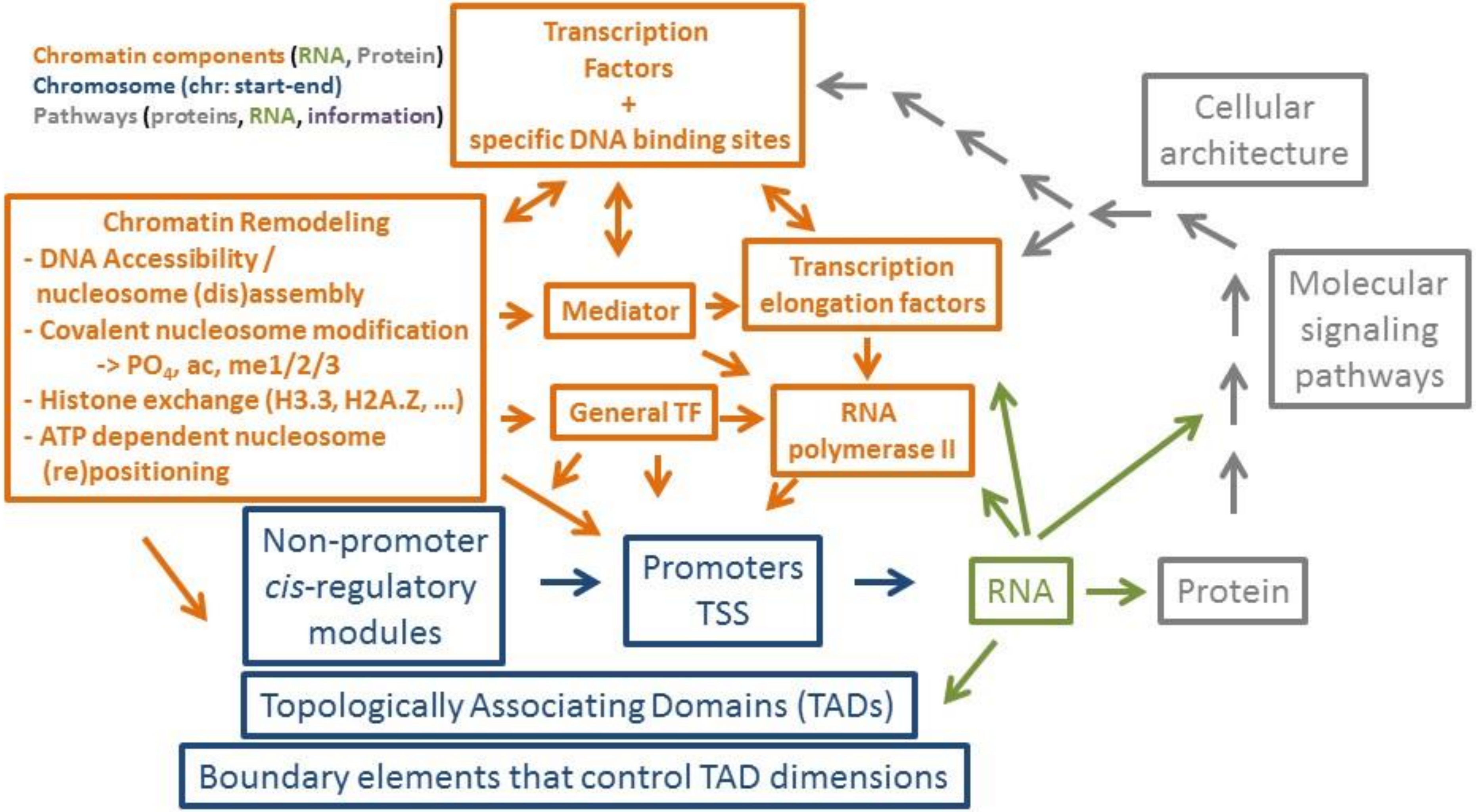
SO term discussion in Lausanne

- Chromosomal Region (*see SO for exact term*):
  - Regulatory Region
    - Is_a: Promoter
      - (has part)Part: Core Promoter
        - Is_a: Core Eukaryotic Promoter
      - Part: TSS
      - Part: CRE
      - Part: TFBS
      - Is_a: Eukaryotic Promoter
        - Part: Core Eukaryotic Promoter
        - Part: Eukaryotic TSS
        - Part: BS for Eukaryotic General (Basal) Transcription Factors
      - Is_a: Bacterial Promoter
      - Is_a: Viral Promoter
    - Is_a: Enhancer
      - Part: TFBS
      - Part: Transcription Initiation Domain
    - Is_a: Silencer
      - Part: TFBS
    - Is_a: Insulator
      - Part: TFBS
    - Is_a: TFBS
    - Is_a: Anchor
      - Part: TFBS
    - Is_a: CRD Boundary
    - Is_a: TAD Boundary
    - Is_a: Transcription Initiation Domain
  - Is_a: DNA Loop
    - Necessary Part: Anchor
  - Is_a: Chromatin Regulatory Region
    - Necessary Part: CRD Boundary
    - Part: DNA Loop
      - Necessary Part Anchor
  - Is_a: Topologically Associated Domain
    - Necessary Part: TAD Boundary
    - Part: Chromatin Regulatory Region
      - Part: CRD Boundary
      - Part: DNA Loop
        - Part Anchor
-

Variant
Gene Regulation High Level Simplified Schema



GO/SO Molecular Functions: signaling to- and by- the epigenome



Chromatin components (RNA, Protein)
Chromosome (chr: start-end)
Pathways (proteins, RNA, information)

Transcription Factors + specific DNA binding sites

Chromatin Remodeling
- DNA Accessibility / nucleosome (dis)assembly
- Covalent nucleosome modification -> PO_4, ac, me1/2/3
- Histone exchange (H3.3, H2A.Z, …)
- ATP dependent nucleosome (re)positioning

Mediator

Transcription elongation factors

General TF

RNA polymerase II

Non-promoter *cis*-regulatory modules

Promoters TSS

RNA

Protein

Cellular architecture

Molecular signaling pathways

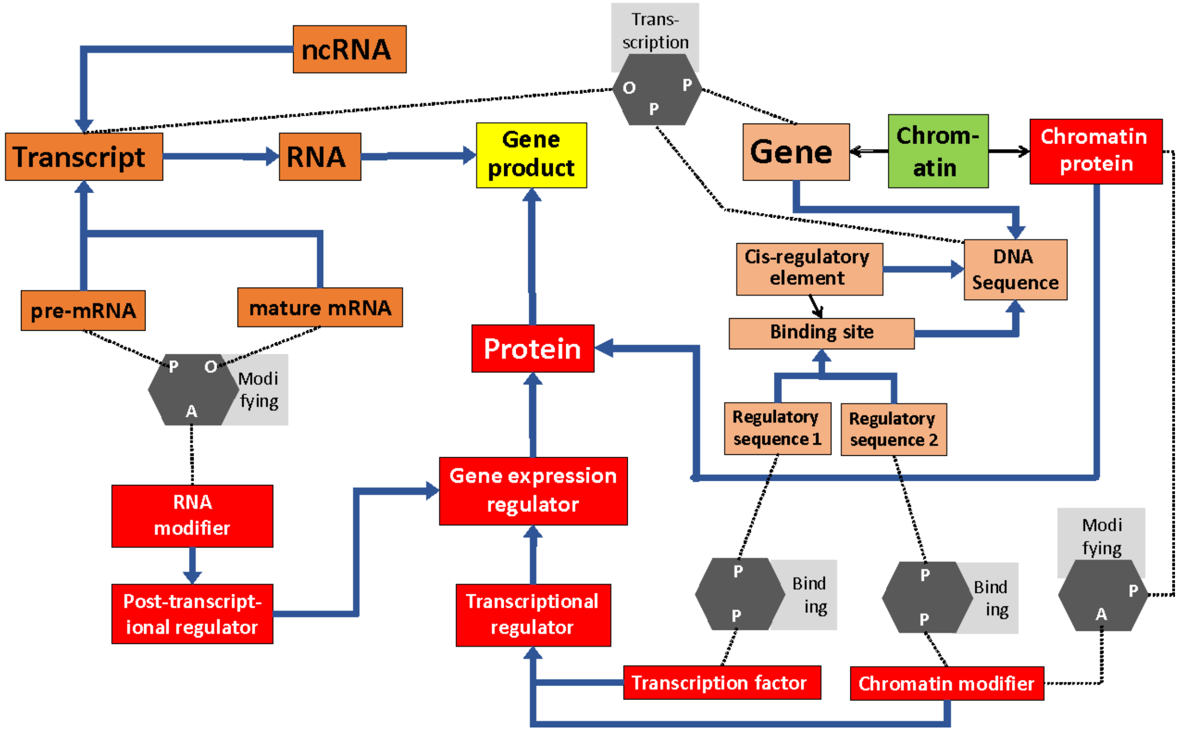
Topologically Associating Domains (TADs)

Boundary elements that control TAD dimensions

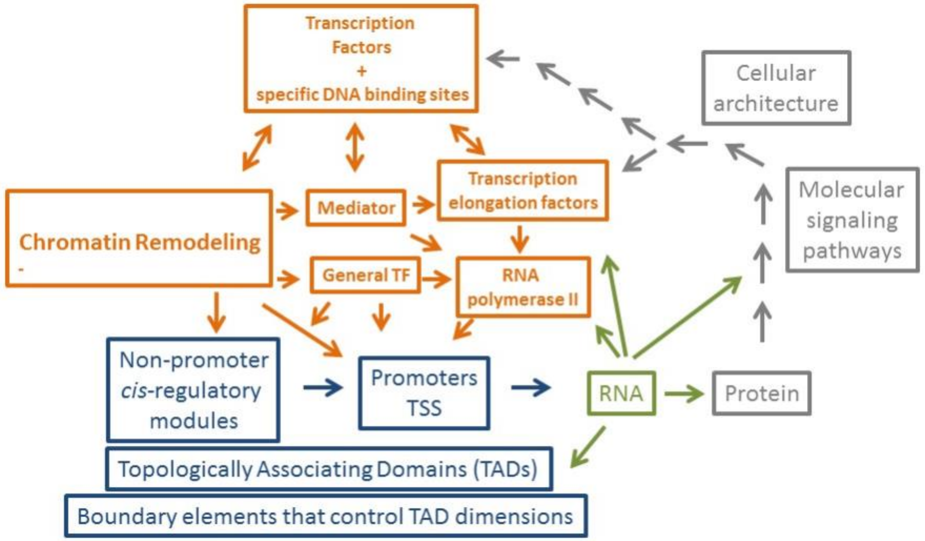
To the genome: Use a list of TFs, chromatin factors and/or non-coding RNAs to discover which parts of the genome are targeted. Identify the transcriptional effects of a list of putative pathway components.

From the genome: Use a list of differentially expressed RNAs/proteins or genomic suites to discover which signaling pathway drives the dynamics. Identify the upstream signals for the putative transcription program.

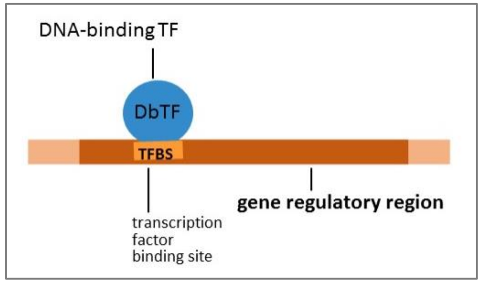
GRO schema

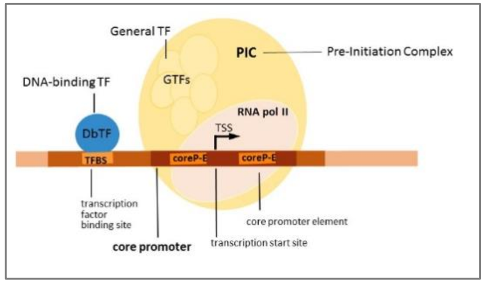


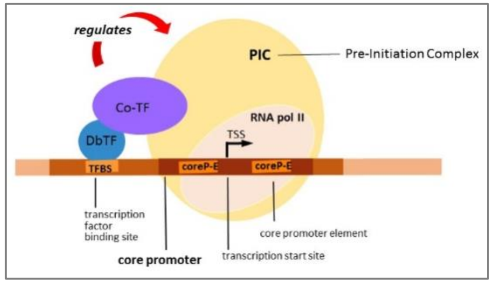
GO/SO schema/flow diagram

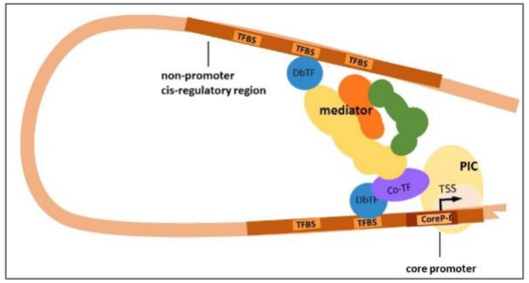


cartoons.....how we envisage the components and interactions





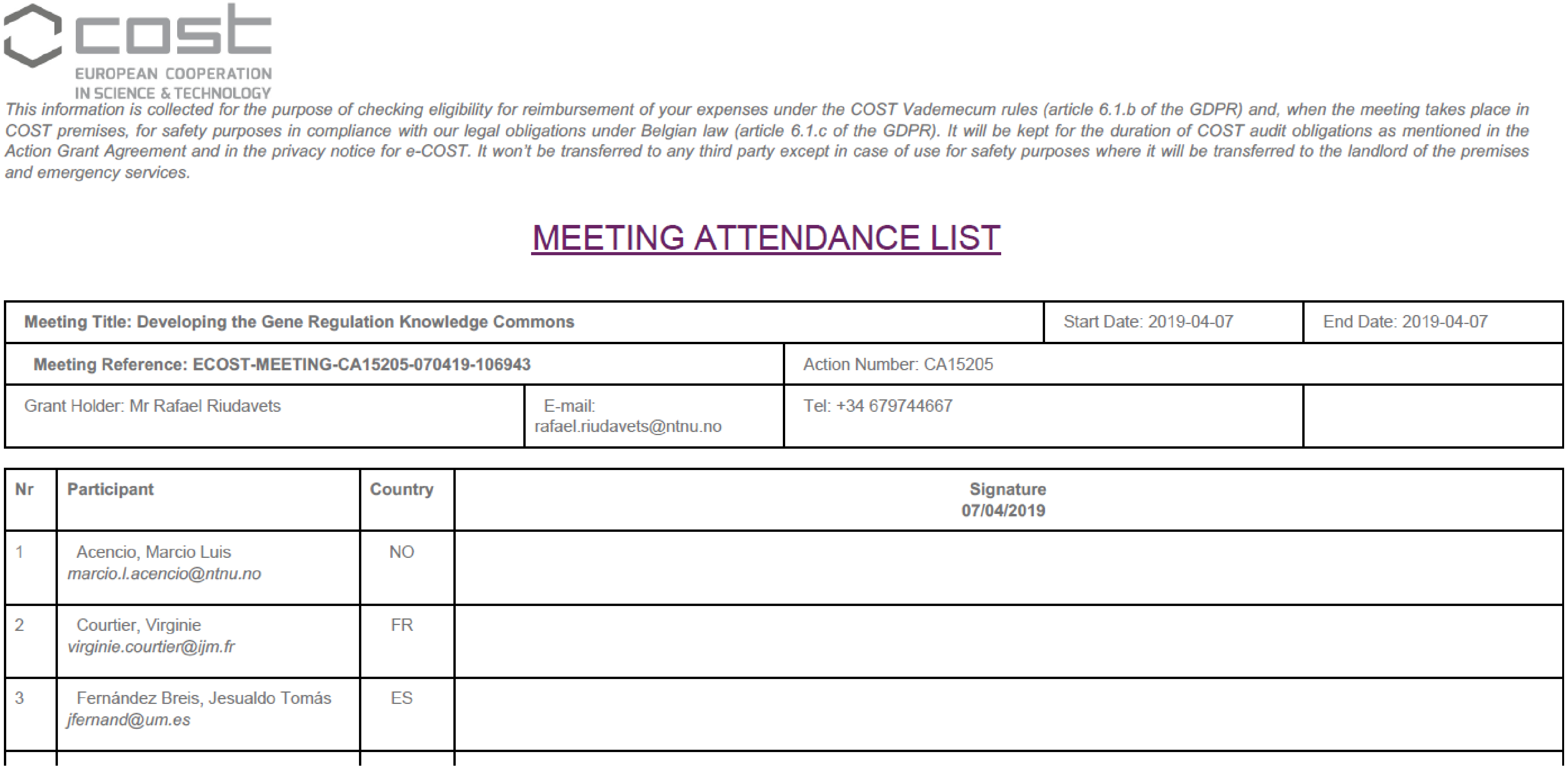




Plans for 2019

- At least 2 Workshops, possibly 3:
  - Ontologies: GO, SO/MSO, GRO, PSI-MI / ECO
  - Curation: Guidelines for specific areas of Gene Regulation
  - Data sharing: interoperability issues of databases and tools
- Training School on ‘interoperability’
- STSMs

Please sign the attendance list!



COST
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

This information is collected for the purpose of checking eligibility for reimbursement of your expenses under the COST Vademecum rules (article 6.1.b of the GDPR) and, when the meeting takes place in COST premises, for safety purposes in compliance with our legal obligations under Belgian law (article 6.1.c of the GDPR). It will be kept for the duration of COST audit obligations as mentioned in the Action Grant Agreement and in the privacy notice for e-COST. It won't be transferred to any third party except in case of use for safety purposes where it will be transferred to the landlord of the premises and emergency services.

MEETING ATTENDANCE LIST

Meeting Title: Developing the Gene Regulation Knowledge Commons		Start Date: 2019-04-07	End Date: 2019-04-07
Meeting Reference: ECOST-MEETING-CA15205-070419-106943		Action Number: CA15205	
Grant Holder: Mr Rafael Riudavets	E-mail: rafael.riudavets@ntnu.no	Tel: +34 679744667	

Nr	Participant	Country	Signature 07/04/2019
1	Acencio, Marcio Luis *marcio.l.acencio@ntnu.no*	NO	
2	Courtier, Virginie *virginie.courtier@ijm.fr*	FR	
3	Fernández Breis, Jesualdo Tomás *jfernand@um.es*	ES	

Contacts:

- Our website: www.greekc.org
- Twitter: @costgreekc
- Grant holder manager: Rafael Riudavets (rafael.riudavets@ntnu.no)
- Action Chair: Martin Kuiper (martin.kuiper@ntnu.no)