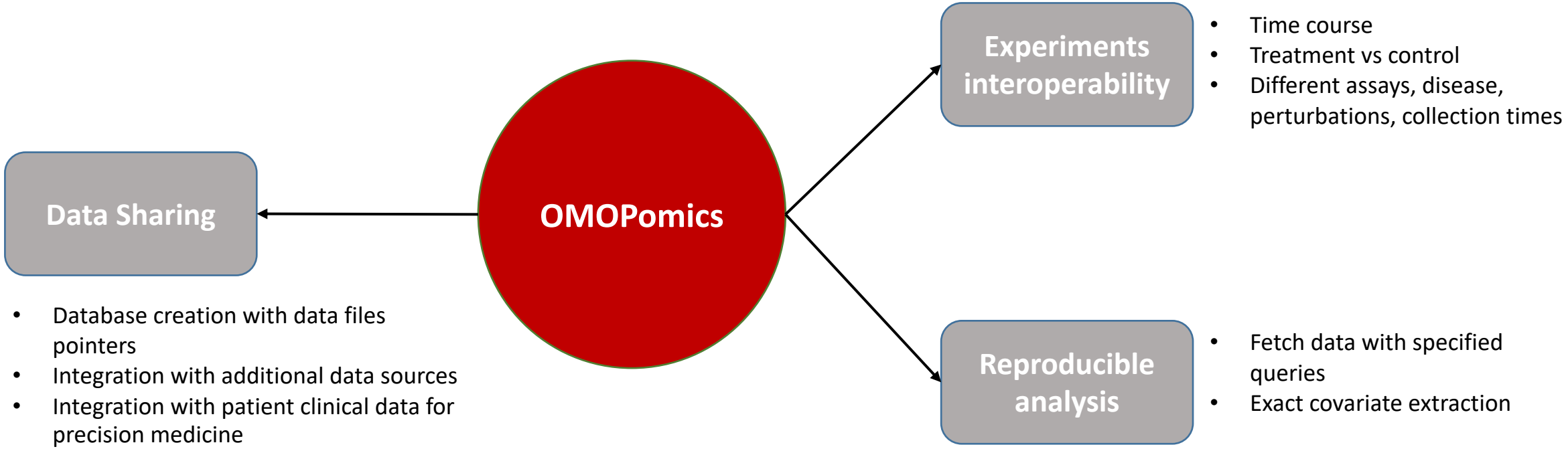
OMOPomics

Day3 update

Nick Giangreco, Andrew Clugtson,
Alex Francette, Anish Chakka, Yash Gokhale

OMOPomics

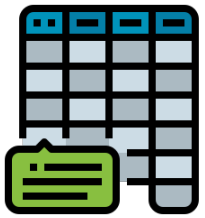
- Observational Health Data Sciences and Informatics (**OHDSI**, pronounced as ***Odyssey***) is a collaborative effort to bring out the value of health data through large-scale analytics.
- OHDSI network formats patient data in OMOP for evidence-based medicine.
- Omics data in OMOP is called **OMOPomics**.



Flowchart

1. Collect and extract data

2. Encode data into OMOP formatted tables



3. Initialize SQL database of OMOP tables

4. Perform SQL queries that output file paths to relevant cohort data

5. Pass file paths to personalized, automated analysis/imaging pipeline

Icons obtained from flaticon.com by authors Wanicon, Freepik, Nhor phai

SPECIMEN_TABLE

Column	Name
specimen_id	S1
specimen_source_value	GSE60682
specimen_type_source_value	CD4+ T cells
person_id	100001

PERSON_TABLE

Column	Name
person_id	100001
gender_concept_id	8507
person_source_value	Donor1
gender_source_value	male

PROVIDER_TABLE

Column	Name
provider_id	P100001
provider_source_value	GEO
provider_type_source_value	GSE60682
person_id	100001

CONDITION_OCCURENCE_TABLE

Column	Name
person_id	100005
condition_occurence_id	C100005
condition_type_value	cutaneous T cell leukemia (CTCL)

ASSAY_OCCURENCE_TABLE

Column	Name
assay_occurrence_id	A1
specimen_source_value	GSE60682
assay_start_date	4:00:00
assay_source_value	ATAC
assay_type_source_value	Sequencing
specimen_id	S1

ASSAY_PARAMETERS_TABLE

Column	Name
specimen_id	S1
assay_parameters_id	AP1
reference_source_value	Homo sapiens
reference_genome_value	hg19

PERTURBATION_TABLE

Column	Name
specimen_id	S1
perturbation_id	Z1
perturbation_source_value	ionomycin
perturbation_type_source_value	activation
perturbation_start_date	2:00:00
perturbation_dose_value_as_number	1
perturbation_dose_unit	ug/mL

ASSAY_OCCURRENCE_DATA _TABLE

Column	Name
assay_occurrence_data_id	AD1
file_source_value	/pylon5/brz3a1p/codeathon/Chroma_T-Cell/Data/GSE60682/GSM1484802_Donor1_TCA4hrs_Rep1.bed.gz
specimen_id	S1

Usage example: ATAC-seq analysis I

- ATAC-seq measures accessible chromatin, reflects epigenomic changes.

- **What epigenomic changes take place following T-cell activation?**

(Buenrostro, Wu, Chang, & Greenleaf, 2015)

Cell Syst. 2015 Jul 29;1(1):51-61.

Individuality and variation of personal regulomes in primary human T cells.

Qu K[1], Zaba LC[1], Giresi PG[1], Li R[1], Longmire M[1], Kim YH[2], Greenleaf WJ[3], Chang HY[1].

Author information

1 Howard Hughes Medical Institute and Program in Epithelial Biology, Stanford University School of Medicine, Stanford, CA 94305.
2 Department of Dermatology, Stanford University School of Medicine, Stanford, CA 94305.
3 Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305.

Abstract

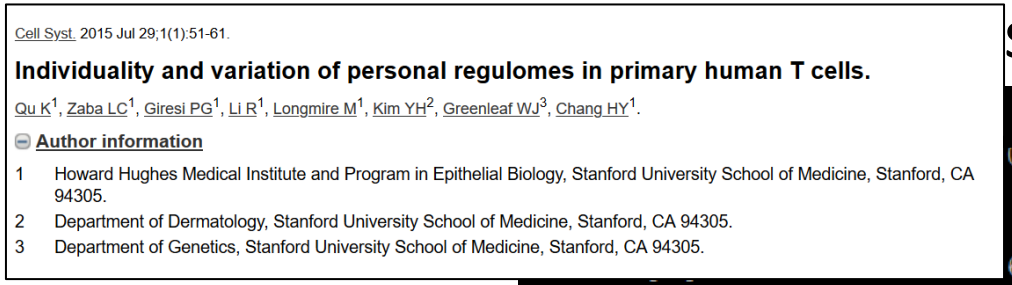
Here we survey variation and dynamics of active regulatory elements genome-wide using longitudinal samples from human individuals. We applied Assay of Transposase Accessible Chromatin with sequencing (ATAC-seq) to map chromatin accessibility in primary CD4+ T cells isolated from standard blood draws of 12 healthy volunteers over time, from cancer patients, and during T cell activation. Over 4,000 predicted regulatory elements (7.2%) showed reproducible variation in accessibility between individuals. Gender was the most significant attributable source of variation. ATAC-seq revealed previously undescribed elements that escape X chromosome inactivation and predicted gender-specific gene regulatory networks across autosomes, which coordinately affect genes with immune function. Noisy regulatory elements with personal variation in accessibility are significantly enriched for autoimmune disease loci. Over one third of regulome variation lacked genetic variation in cis, suggesting contributions from environmental or epigenetic factors. These results refine concepts of human individuality and provide a foundational reference for comparing disease-associated regulomes.

PMID: 26251845 PMCID: PMC4522940 DOI: 10.1016/j.cels.2015.06.003

Usage example: ATAC-seq analysis II

1. Use search query to pull accessibility time course files measured at 0, 1, 2, and 4 hours post activation.

2. Read peak regions and counts into R.

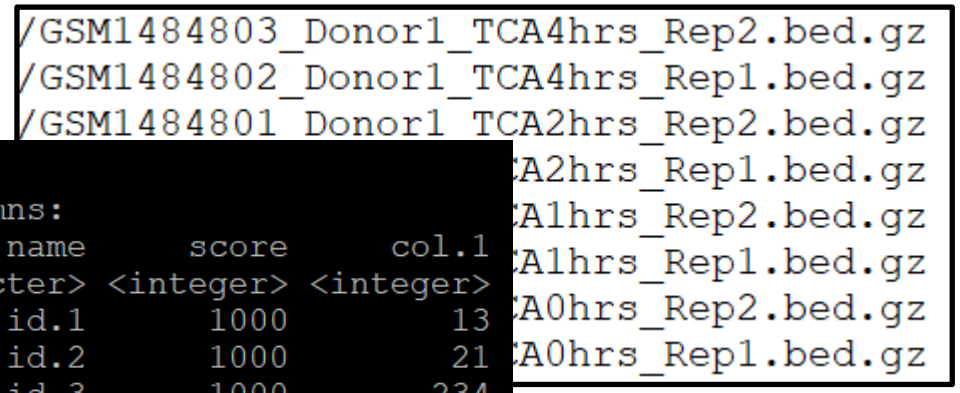

Cell Syst. 2015 Jul 29;1(1):51-61.

Individuality and variation of personal regulomes in primary human T cells.

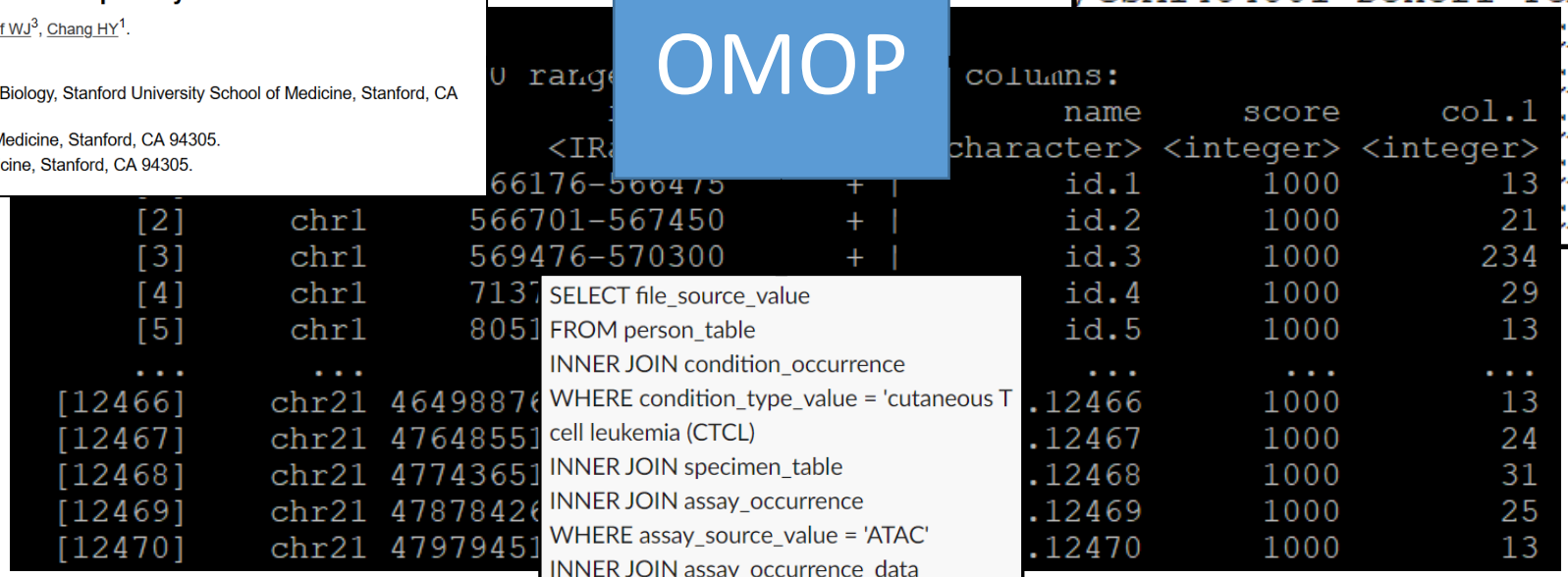
Qu K[1], Zaba LC[1], Giresi PG[1], Li R[1], Longmire M[1], Kim YH[2], Greenleaf WJ[3], Chang HY[1].

Author information

1 Howard Hughes Medical Institute and Program in Epithelial Biology, Stanford University School of Medicine, Stanford, CA 94305.
2 Department of Dermatology, Stanford University School of Medicine, Stanford, CA 94305.
3 Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305.



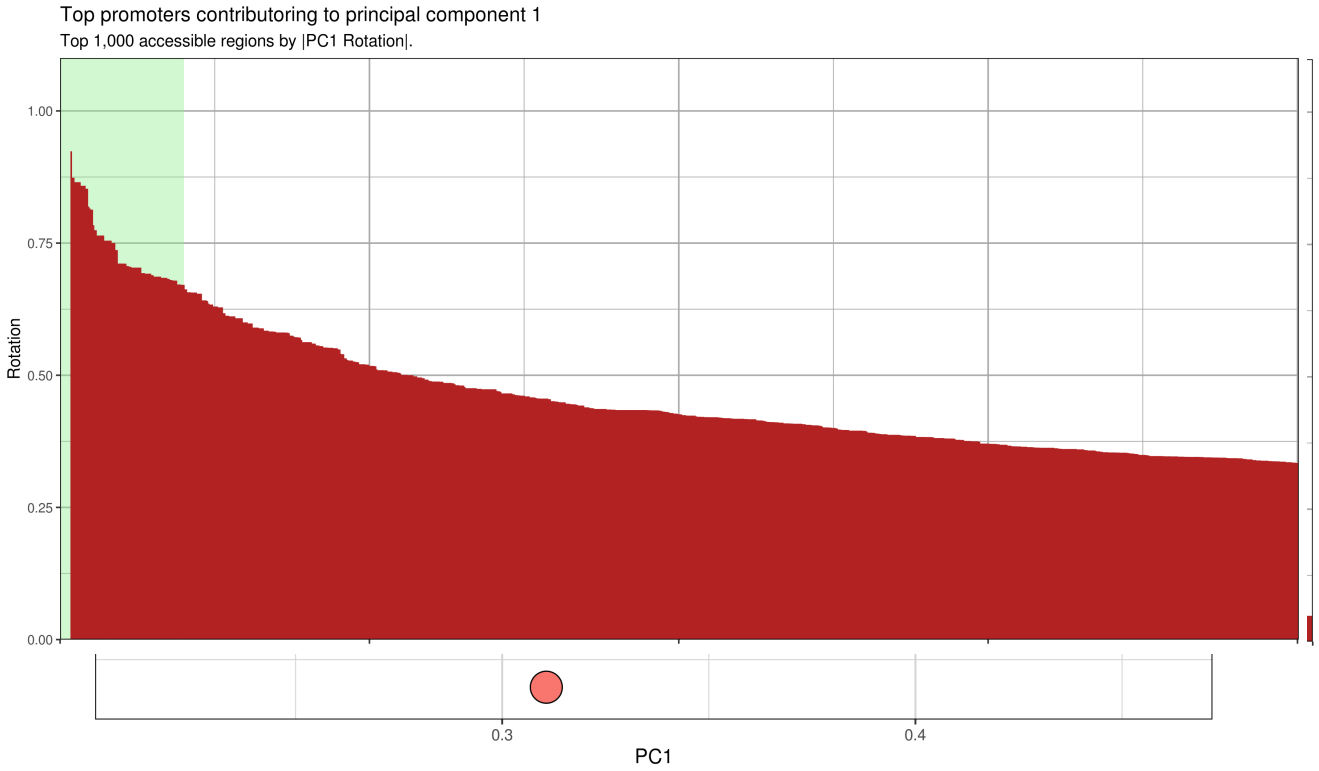
OMOP




```
SELECT file_source_value
FROM person_table
INNER JOIN condition_occurrence
WHERE condition_type_value = 'cutaneous T
cell leukemia (CTCL)
INNER JOIN specimen_table
INNER JOIN assay_occurrence
WHERE assay_source_value = 'ATAC'
INNER JOIN assay_occurrence_data
```

Usage example: ATAC-seq analysis III

1. How many accessible regions are found in each sample?
2. What transcript promoters open/close?
3. How similar is the epigenomic profile betwe samples?
   - …Promoters only?
4. Which genomic regions drive PC1?
   - …Promoters only?
5. Etc.



Main Conclusions

- Extended existing clinical data infrastructure to promote interoperable and reproducible data storage/analysis of biological data
- Promoted integration of patient clinical and molecular data
- Enumerated experimental processing to improve understanding and downstream analysis

OMOPomics post-codeathon

- Communicating through Slack channel and GitHub issues
- Contacting OHDSI project manager (Maura)
  - Reserve presentation on weekly conference call
  - On the look out for OHDSI 2020 conference call-for-papers
- Posting OMOPomics on OHDSI forum for feedback and collaborations
- Open Science Foundation project created for OMOPomics
- Writing manuscript in Google documents
- Adding data from other projects
  - Data from SVAI?
  - GTEx?
  - TCGA?