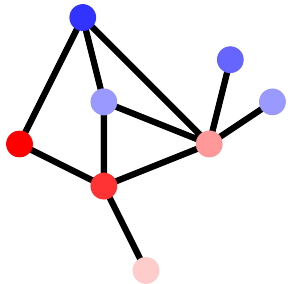
PCAWG-5:
Pathway and network analysis update



Ben Raphael
Josh Stuart
On behalf of PCAWG-5

PCAWG-2,5,9,14 Joint Call
December 4, 2017

Background and Overview

Data:

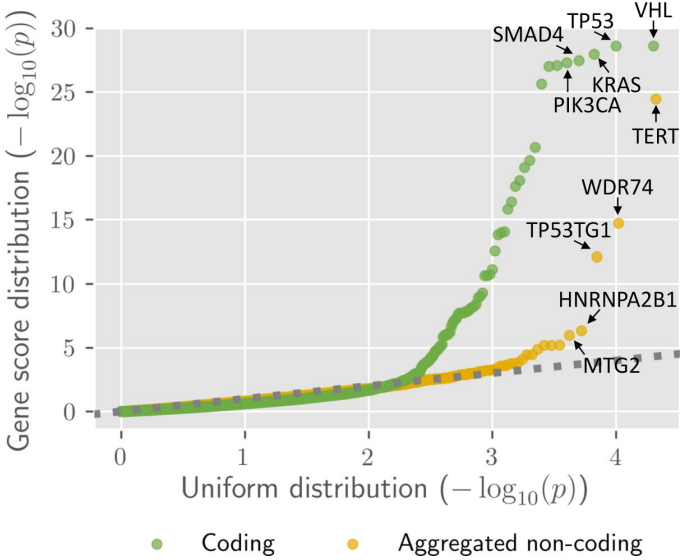
- Integrated driver scores from PCAWG-2-5-9-14 (syn8494939)
- Focused analysis on Pancan-no-skin-melanoma-lymph cohort
- 63 genes with recurrent (q < 0.1) **coding** mutations
- 6 genes with recurrent (q < 0.1) **non-coding** mutations (promoter core, UTRs, or enhancer):
    - HES1 promoter core, MTG2 5' UTR, TERT promoter core, TOB1 3' UTR, TP53TG1 enhancer, WDR74 promoter core

Motivation:

- Few identified non-coding drivers
- Pathway/network analyses helpful for analyzing rare coding mutations in earlier studies

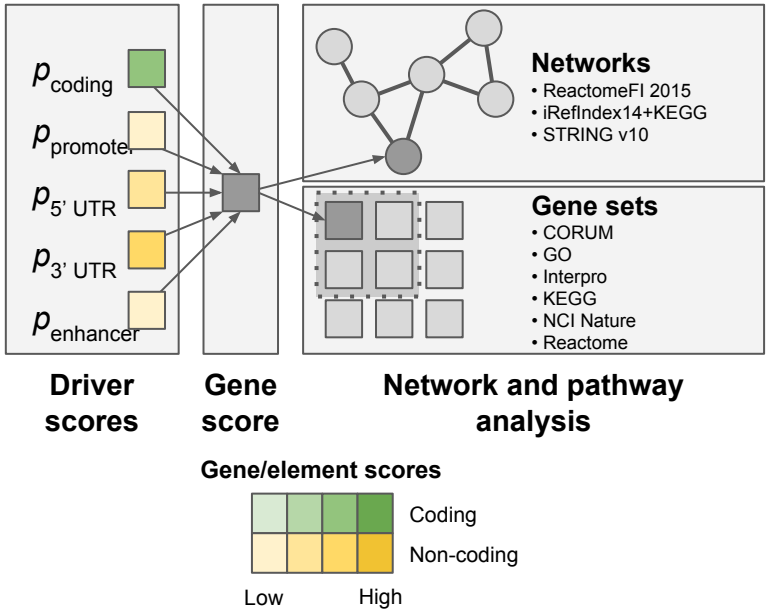
Questions:

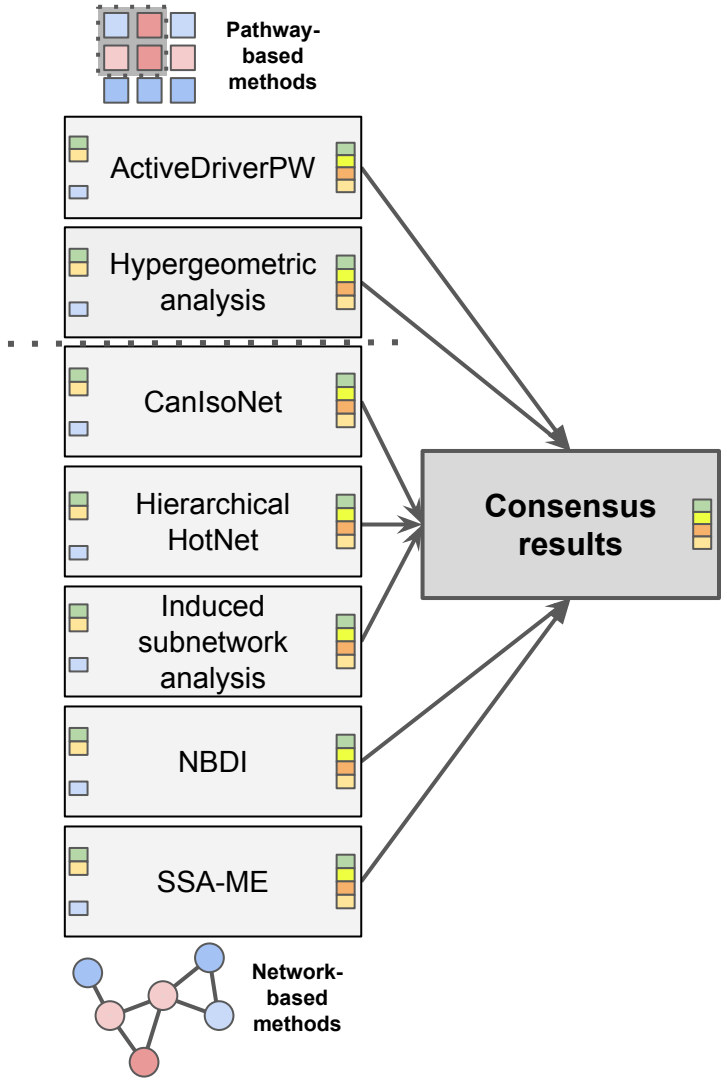
1. Can pathway/network methods identify additional coding and non-coding drivers? (Dig deeper into long tail?)
2. Do non-coding drivers cluster in pathways/subnetworks?
3. Do non-coding drivers cluster in *same* pathways/subnetworks as coding drivers?



Analysis overview

1. Create **gene-level** scores from coding and non-coding elements.
2. Apply pathway/network methods to coding and non-coding scores, separately and combined.
3. Form consensus results across methods.
4. Analyze contributions of coding and non-coding mutations to enriched pathways/networks.



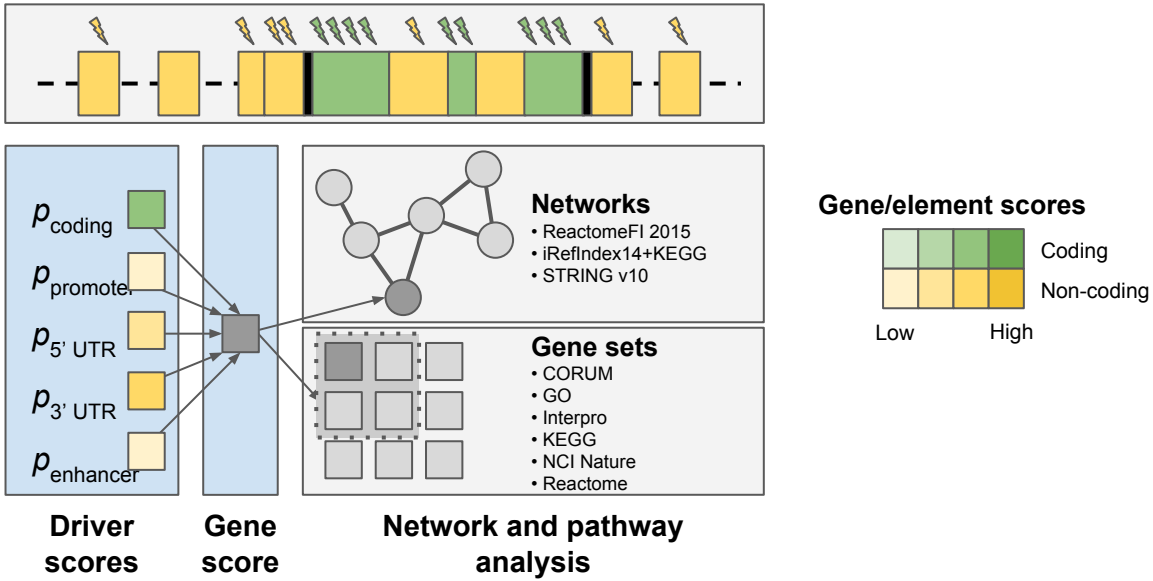


Create gene-level scores from coding and non-coding element scores

Pathway/network databases typically describe protein/gene interactions, gene-level scores needed.

Define coding (***C***), non-coding (***N***), and combined coding and non-coding (***C+N***) scores for each gene:

1. $p_C = p_{coding}$
2. $p_N = \text{fisher}(\min(p_{promoter}, p_{5' UTR}), p_{3' UTR}, p_{enhancer})$
3. $p_{C+N} = \text{fisher}(p_{coding}, \min(p_{promoter}, p_{5' UTR}), p_{3' UTR}, p_{enhancer})$



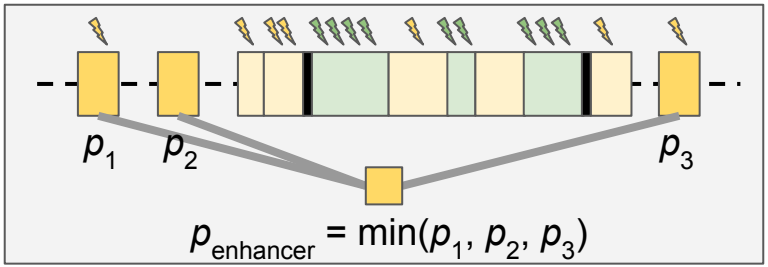
Create gene-level scores for enhancers

Pathway/network databases typically describe protein/gene interactions, gene-level scores needed.

Define gene-level enhancer scores:

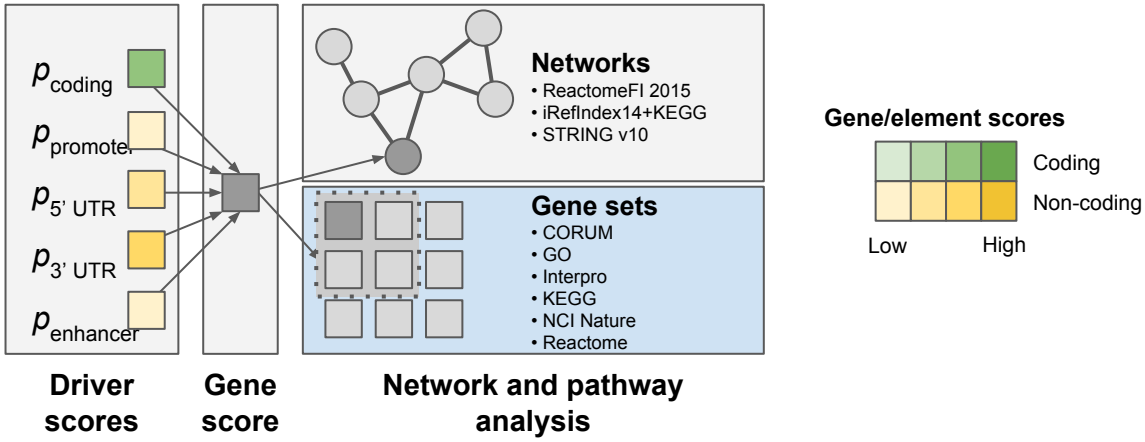
1. Use enhancer scores (syn8494939).
2. Use enhancers with ≤ 5 gene targets from enhancer-gene targets (syn7188184).
3. For each gene, define $p_{enhancer}$ as **minimum score** of enhancers targeting gene.

We also performed analysis without enhancer scores.



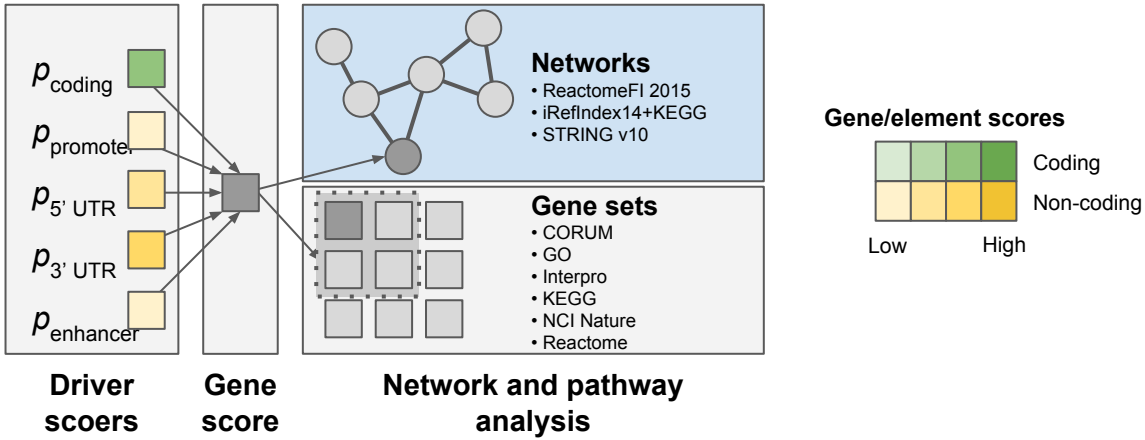
Pathway analyses

1. **ActiveDriverPW (Jüri Reimand, OICR)**: Evaluates the enrichment of mutations in functionally related sets of genes using driver scores.
2. **Hypergeometric analysis (Miguel Vazquez, CNIO)**: Identifies enriched pathways (hypergeometric test *q*-value $q < 0.05$) containing genes with driver score *p*-values $p < 0.1$.

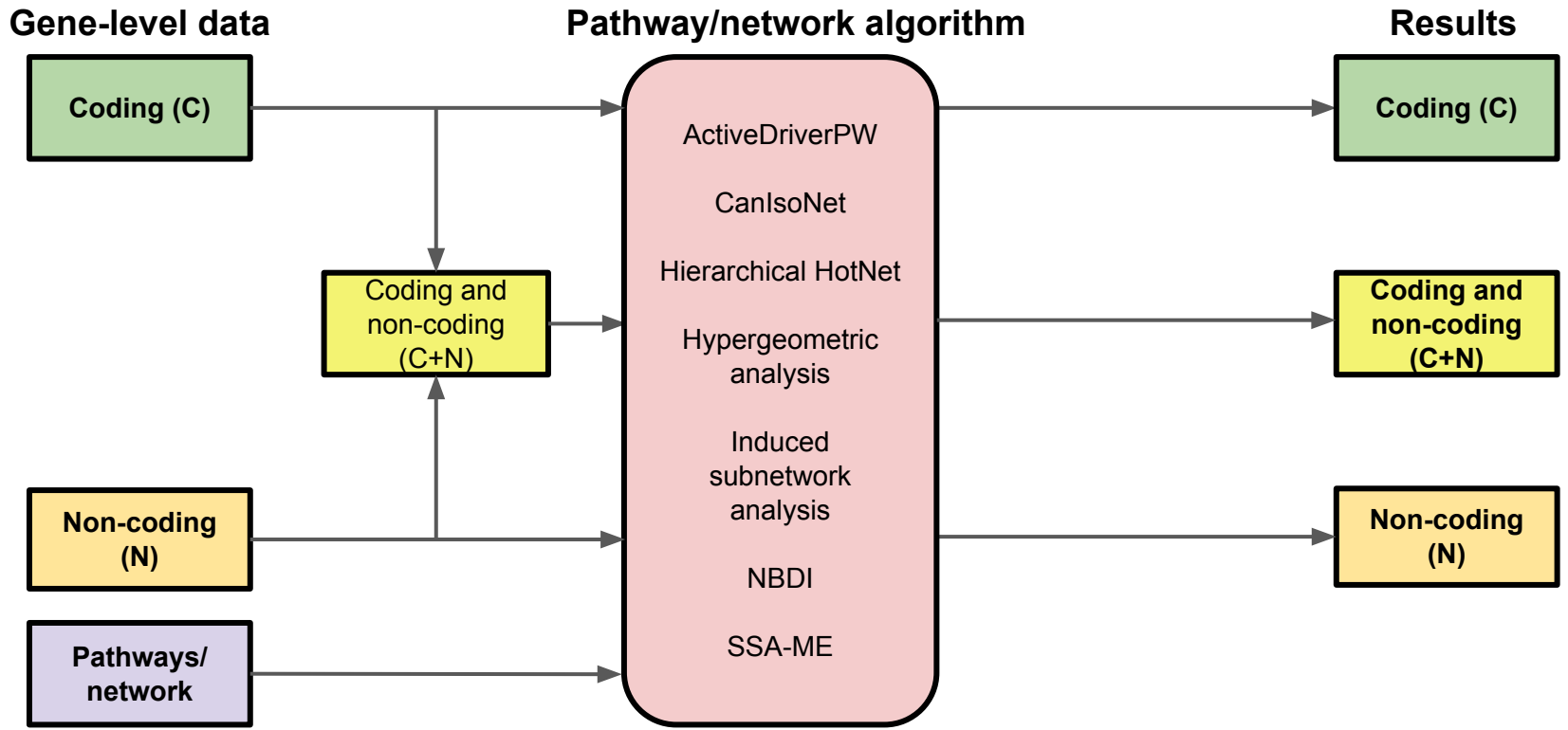


Network analyses

1. **CanIsoNet (Abdullah Kahraman, UZH):** Identifies regions of a PPI network whose interactions are considered to be disrupted by alternative isoforms.
2. **Hierarchical HotNet (Matthew Reyna, Princeton):** Combines driver scores and network topology to construct hierarchy of topologically close and significantly mutated gene sets.
3. **Induced subnetwork analysis (Matthew Reyna, Princeton):** Finds *connected* subnetworks induced by genes with driver scores exceeding a statistically determined threshold.
4. **NBDI (Lieven Verbeke, Ghent):** Uses driver scores with patient mutation *and expression* data to find mutated genes that interact with differentially expressed genes.
5. **SSA-ME (Sergio Pulido-Tamayo, Ghent):** Prioritizes genes based on likelihood of belong to high-scoring subnetworks.



Each pathway/network method applied to three datasets: ***C***, ***N***, ***C + N***

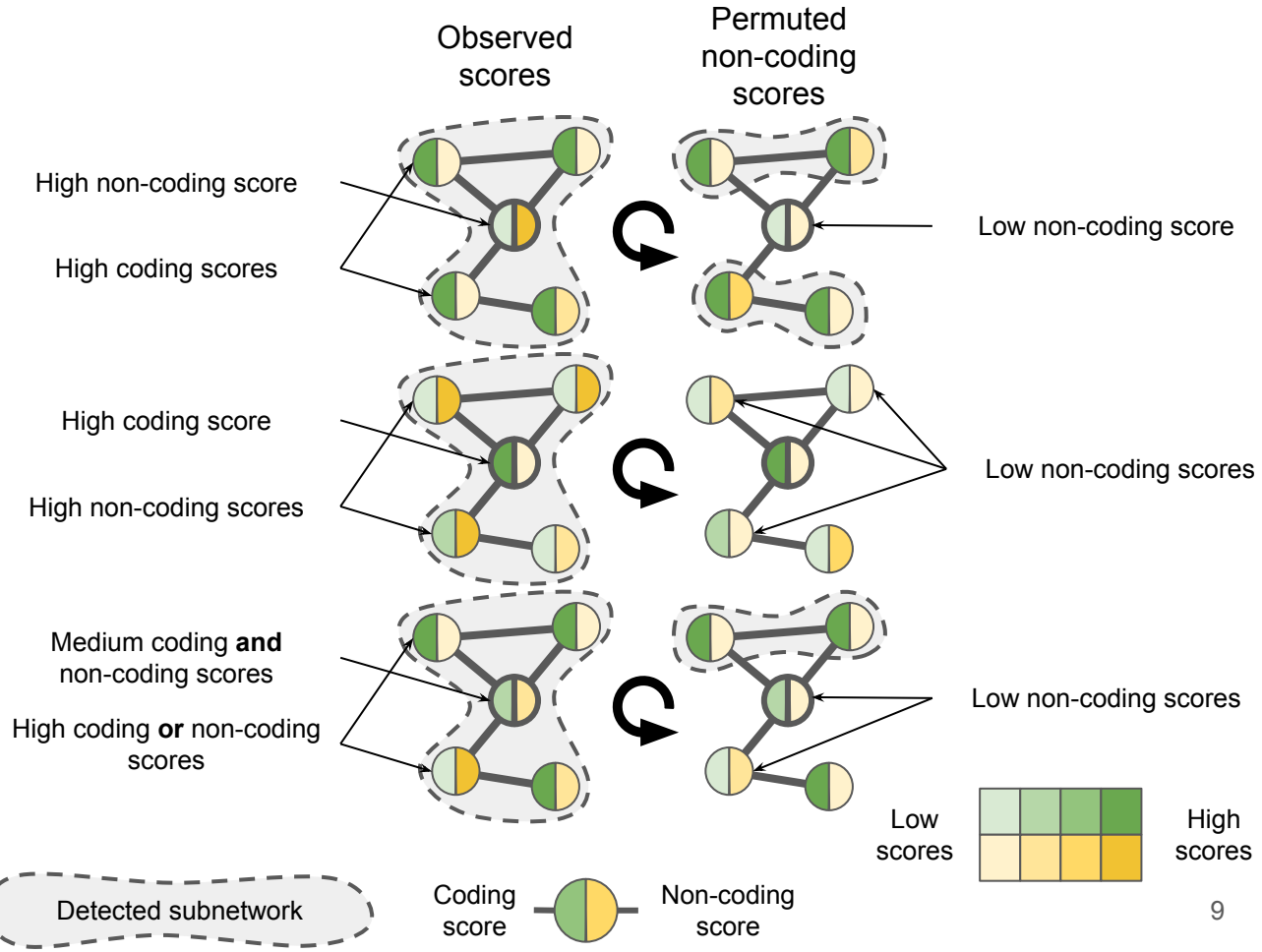


Observation: Stronger coding signal dominant signal in combined coding and non-coding analysis?

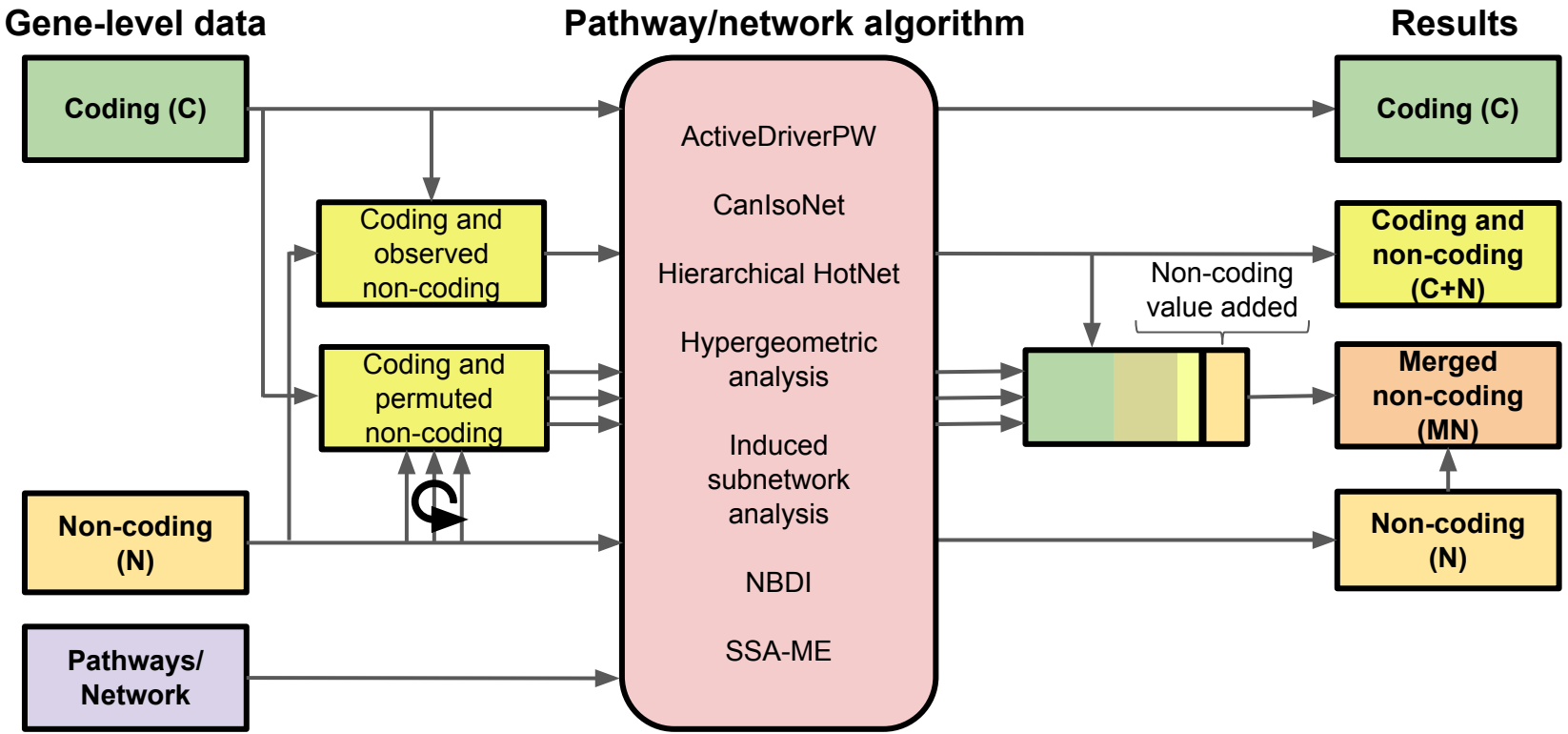
Can we identify genes with strong non-coding contribution in combined analysis?

Non-coding “value-added” procedure

- **Motivation**: Want to increase signal by combining coding and non-coding scores; however **coding scores dominate combined signal**
- **Procedure**: Run pathway network analysis on permuted data: **fixing** coding scores and permuting non-coding scores.
- **Results: Define** gene as a **non-coding value-added** provided gene appears in method’s C+N results in real data, but infrequently ($p < 0.1$) with permuted data.
- Procedure identifies genes with **strong contribution from non-coding scores** while **leveraging coding scores**.



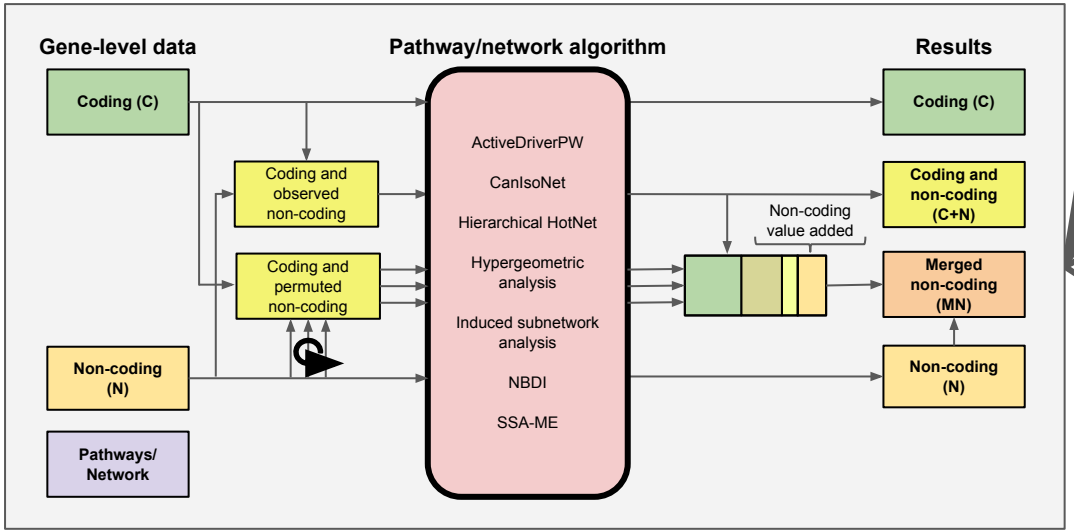
Apply pathway/network methods to coding and non-coding data

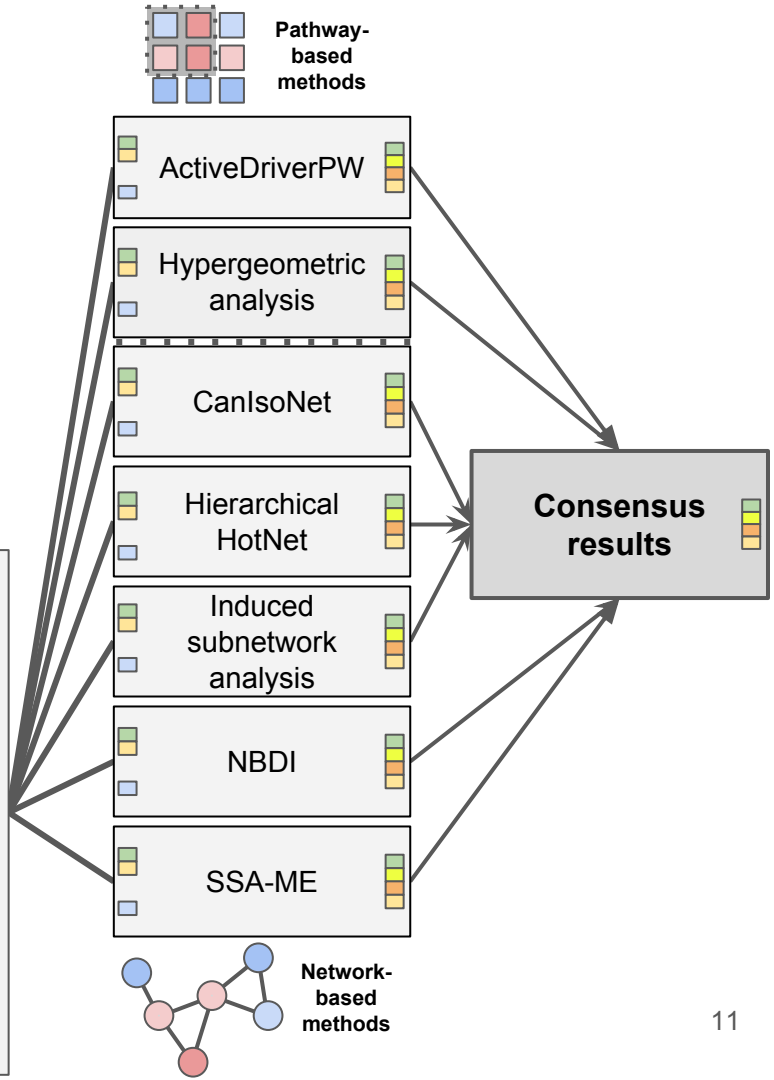


Merged non-coding (MN) results are union of the non-coding (N) and non-coding value-added results.

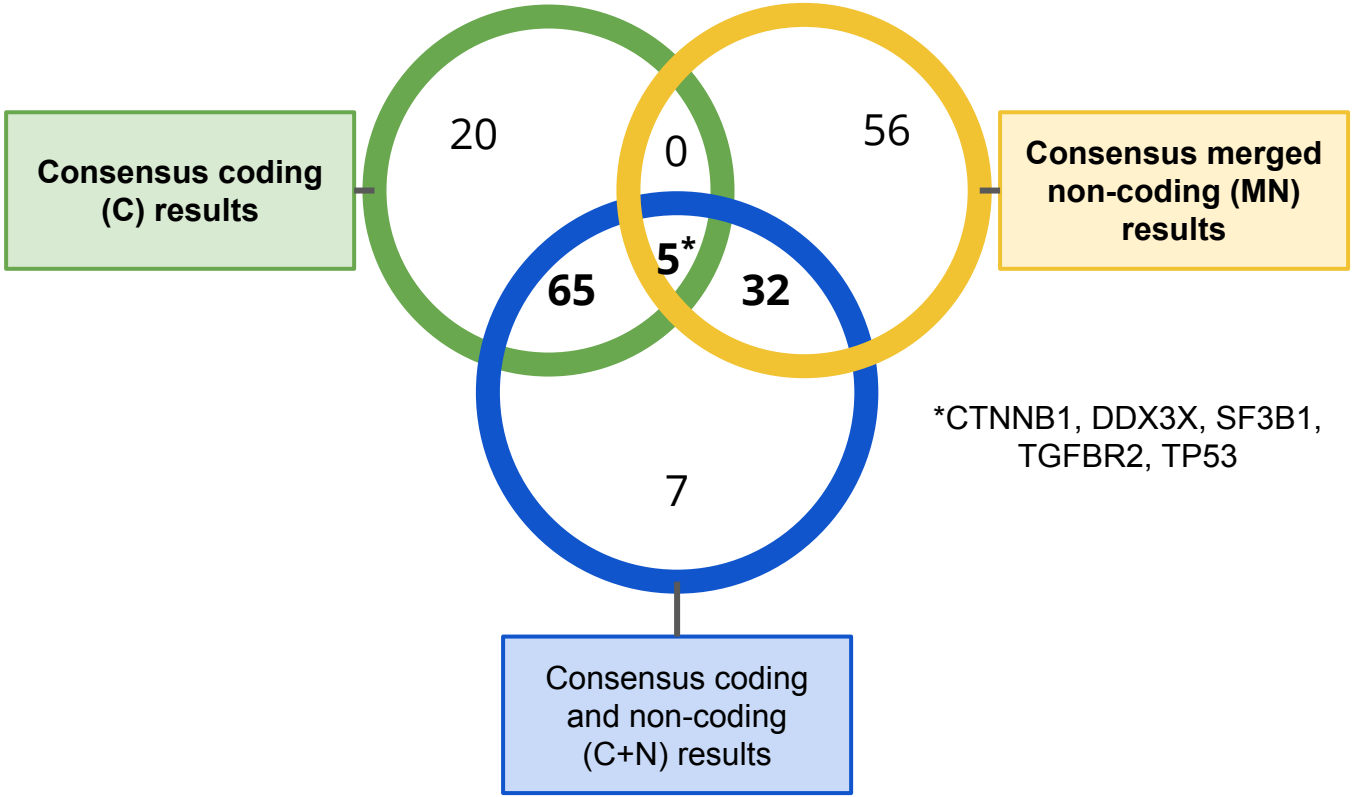
Form consensus results across methods

- Consensus procedure identifies genes by a majority (≥4/7) of methods.
- Perform consensus across methods for each set of results:
  - Coding (C)
  - Non-coding (N)
  - Coding and non-coding (C+N)
  - Merged non-coding (MN)
- Focus on consensus **coding (C)** and **merged non-coding (MN)** results.





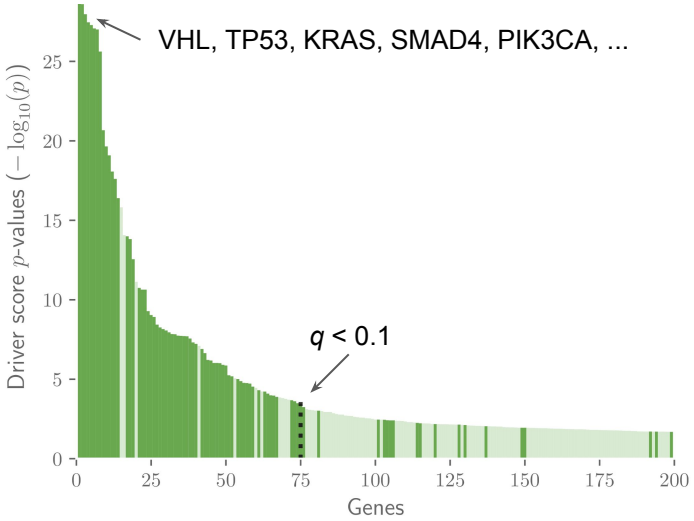
Overlap of consensus pathway/network analysis results

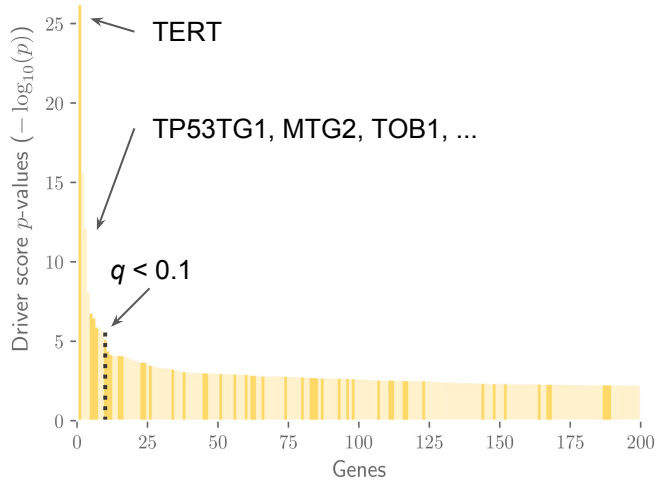


*CTNNB1, DDX3X, SF3B1, TGFBR2, TP53

Comparison with individual gene/element ranking

- **Question 1:** Can pathway/network methods identify additional coding and non-coding drivers? (Dig deeper into long tail?)
- Yes:
  - 31/87 consensus coding genes have driver scores with $q > 0.1$.
  - 90/93 consensus merged non-coding genes have driver scores with $q > 0.1$.



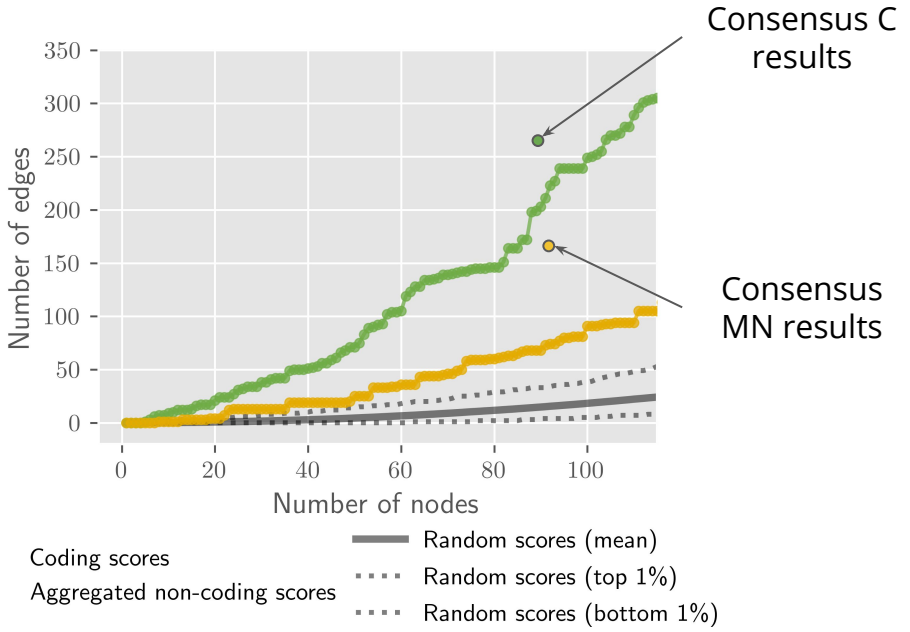


Driver score distributions for coding (**left**) and non-coding (promoter core, UTRs, and enhancer elements; **right**); **dark bars** indicate **consensus genes** and **light bars** indicate **non-consensus genes**.

Clustering of non-coding genes on pathways/networks

Question 2: Do non-coding mutations cluster in pathways/subnetworks?

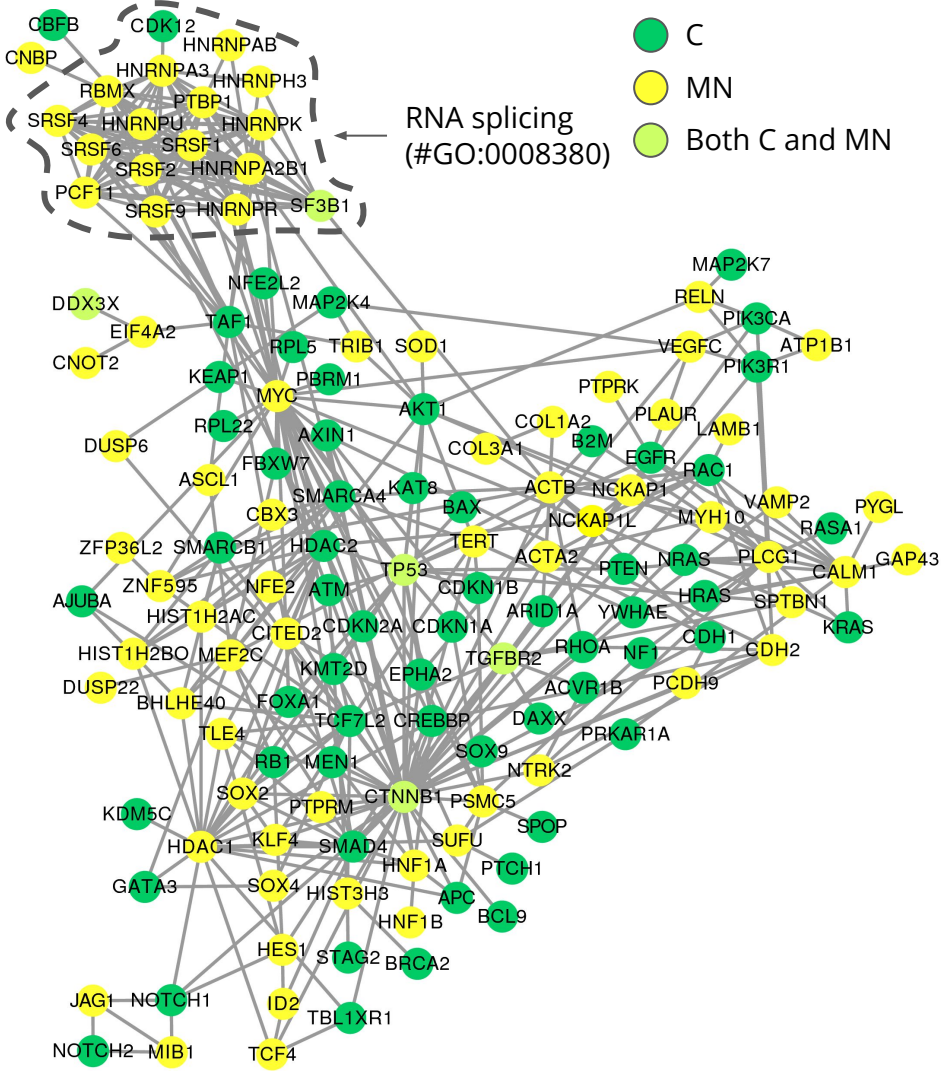
- Yes:
  - Consensus genes form large connected components in ReactomeFI 2015 network:
    - 75 genes (P < 1e-6) for consensus coding results
    - 61 genes, (P < 1e-6) for consensus merged non-coding results
  - Consensus genes are enriched in GO and Reactome pathway databases:
    - 63 pathways with hypergeometric P < 1e-6 for consensus C results
    - 13 pathways with P < 1e-6 for consensus merged non-coding results



Network view of consensus results

Question 3: Do non-coding mutations cluster in same pathways/subnetworks as coding mutations?

- Yes and no.
  - Some interactions primarily between consensus coding results
  - Some interactions primarily between merged non-coding results.
  - Many interactions between consensus coding and consensus merged non-coding results.

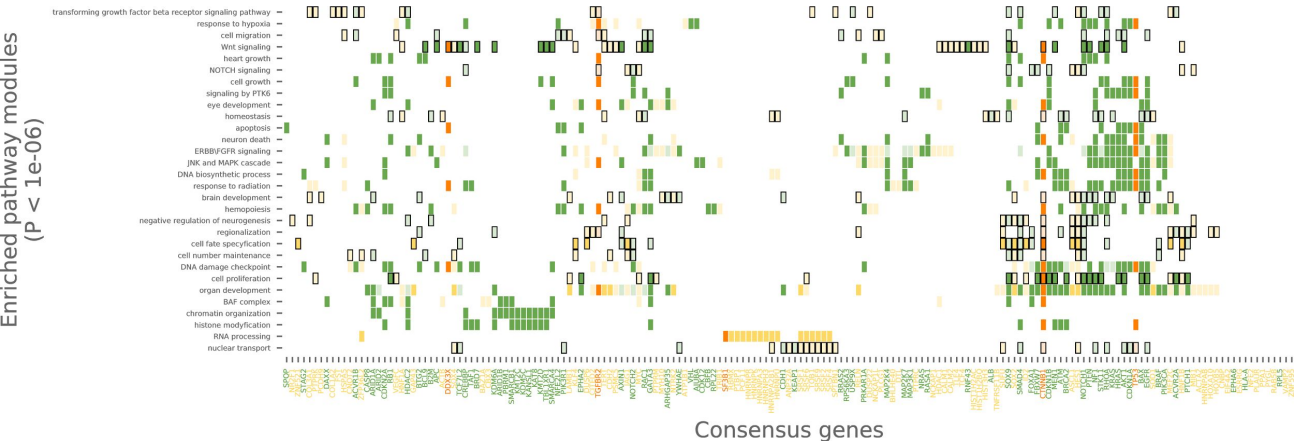


Subnetwork of ReactomeFI 2015 network induced by consensus coding (C) and consensus merged non-coding (MN) results, **excluding** 214 interactions only between genes in coding results

Pathway view of consensus results

Question 3: Do non-coding drivers cluster in same pathways/subnetworks as coding drivers?

- Yes and no.
  - Some enriched pathways primarily due to **consensus coding** results.
  - Some enriched pathways primarily due to c**onsensus merged non-coding results**.
  - Some enriched pathways have **contributions** from **consensus coding and merged non-coding results**.



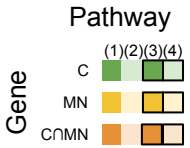
Matrix entries correspond to pathway modules (rows) and a consensus genes (columns).

A filled entry of the matrix indicates

- ≥ 1 pathway in module is enriched for consensus results and
- gene belongs to ≥ 1 of enriched pathways in module.

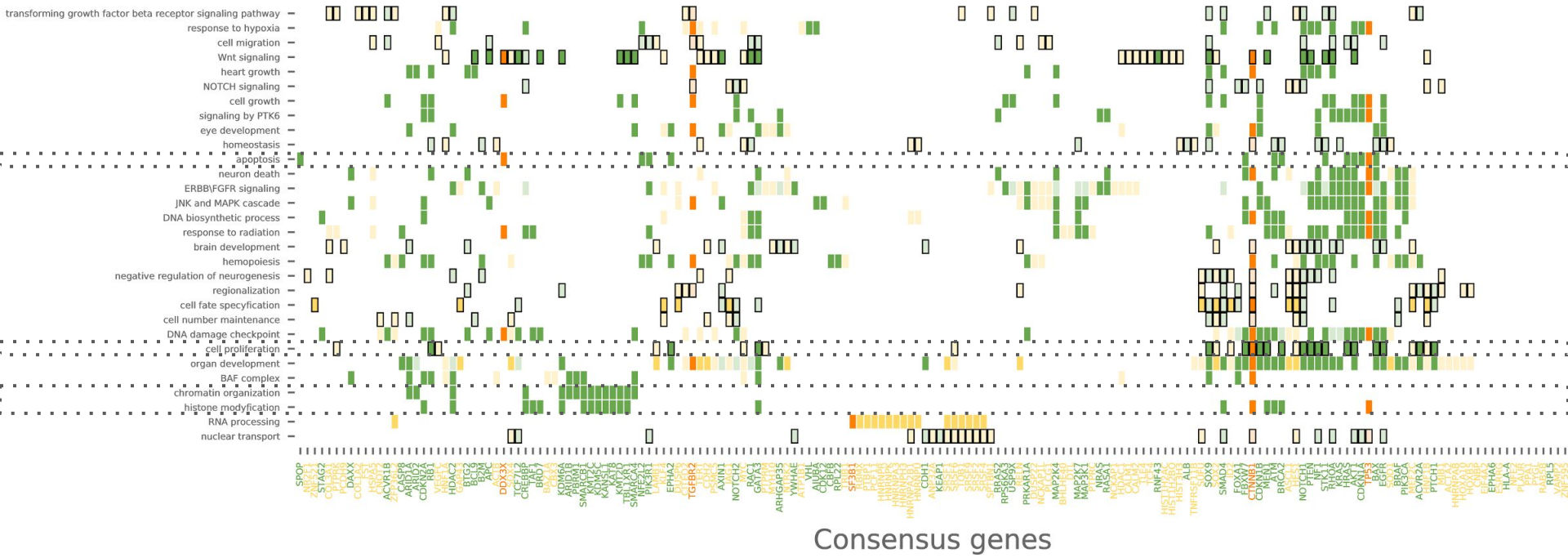
Details:

- **Color**: consensus C or consensus MN gene
- **Shading**: enrichment
- **Border**: more enriched in C ∪ MN than separately



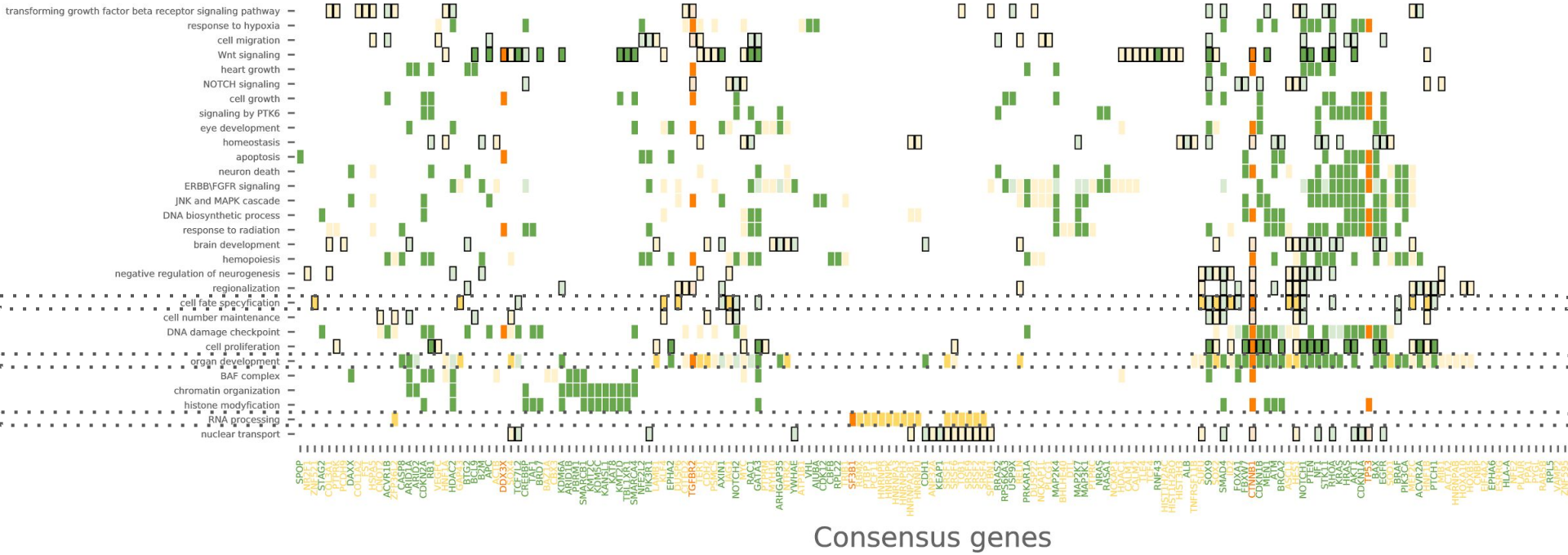
(1) Enriched.
(2) Present, not enriched.
(3) Enriched, C∪MN more significantly enriched.
(4) Present, not enriched, C∪MN enriched.

Pathway view of consensus results: primarily **coding** contributions



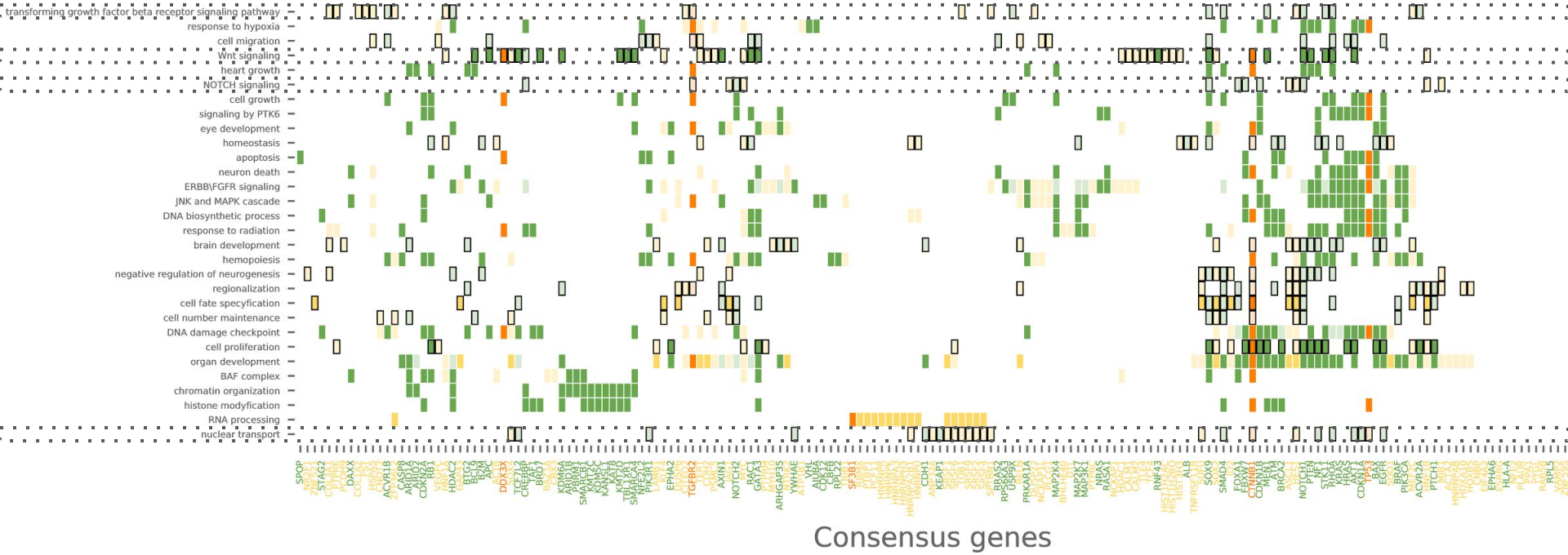
- Apoptosis…
- Cell proliferation...
- Chromatin organization and histone modification...

Pathway view of consensus results: primarily **non-coding** contributions



- Splicing...
- Cell date specification…
- Some development...

Pathway view of consensus results: **both** coding and non-coding contributions



- Signaling pathways (Wnt, NOTCH, growth factor…)
- Nuclear transport…
- Transcription factors...

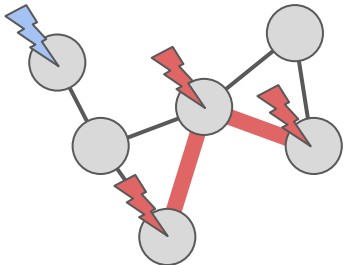
Do driver mutations in a patient occur on pairs of interacting genes?

Data:

- 9,947 **patient-centric driver mutations** in ICGC PCAWG data (see Supplemental Table 3 of bioRxiv paper in syn11050201)
- Four interaction networks: BioGRID HC, HINT+HI2014, iRefIndex14+KEGG, ReactomeFI 2015

Questions:

1. For **each patient**, are there **significantly more** or **significantly fewer interactions** between driver mutations than expected by chance?
2. For **each cohort**, are there **significantly more** or **significantly fewer interactions** between driver mutations than expected by chance?
3. For **each pair of interacting genes**, are both genes affected by driver mutations in **significantly more** or **significantly fewer patients** in a cohort than expected by chance?



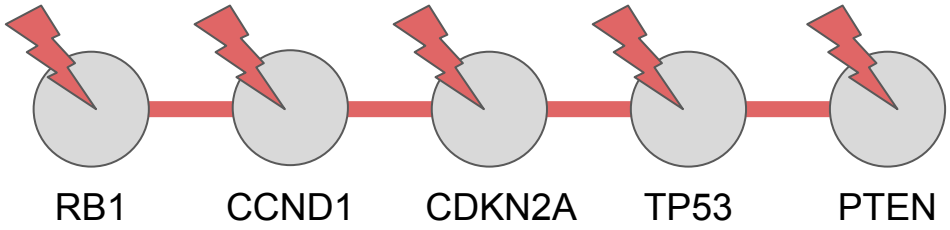
Example

- 4 genes with driver mutations
- 2 interactions between genes with driver mutations

Do driver mutations in a patient occur on pairs of interacting genes?

Example:

- A Liver-HCC patient has 5 driver mutations in BioGrid HC: CCND1, CDKN2A, PTEN, RB1, and TP53.
- Subgraph of BioGrid HC induced by these genes has 4 interactions: CCND1-CDKN2A, CCND1-RB1, CDKN2A-TP53, PTEN-TP53.
- Are there more or fewer interactions in this patient's induced subgraph than expected?
- Are there more or fewer edges in this cohort's induced subgraphs than expected?
- Do any of the interactions, e.g., PTEN-TP53, occur more or less frequently in a cohort than expected?



Summary of patient-specific driver mutations on networks

Network	None (not restricted to network genes)	BioGRID HC	HINT+HI2014	iRefIndex14 + KEGG	ReactomeFI 2015
Number of genes/loci with **driver** mutations	550	396	476	493	457
Number of patients with **driver** mutations	2068	2023	2052	2056	2044
Number of **driver** mutations	9947	7977	8648	8800	8505

Table: Summary of patient-centric driver mutation data on **25 cohorts** with ≥ 10 samples with driver mutations and ≥ 2 driver mutations, on average, per sample in all networks.

Statistical test

We compared the observed results against 10,000 permutations of the binary mutation matrix for each cohort, **preserving** the **mutation frequency** of **each gene** within the cohort and **each patient** in the cohort.

- For a fixed patient *s*, let X_s = number of interactions between driver mutations in *s*. We compute $p_s^+ = \Pr(X_s \geq x_s)$ and $p_s^- = \Pr(X_s \leq x_s)$, where x_s is observed value.

- For a fixed cohort *C*, let $X_C = \sum_{s \in C} X_s$ = number of interactions between driver mutations across patients in *C*. We compute $p_C^+ = \Pr(X_C \geq x_C)$ and $p_C^- = \Pr(X_C \leq x_C)$, where x_C is observed value.

- For fixed cohort *C* and edge *e* = (*u*, *v*), let $X_{C,e}$ = number of patients in *C* with driver mutations in both *u* and *v*. We compute $p_{C,e}^+ = \Pr(X_{C,e} \geq x_e)$ and $p_{C,e}^- = \Pr(X_{C,e} \leq x_{C,e})$.

Table of *p*-values for fewer or more interactions than expected in each cohort

	A	B	C	D	E	F	G	H	I
1	**Cohort**	**p-value for fewer edges than expected in BioGRID HC**	**p-value for fewer edges than expected in HINT+HI2014**	**p-value for fewer edges than expected in iRefIndex14+KEGG**	**p-value for fewer edges than expected in ReactomeFI 2015**	**p-value for more edges than expected in BioGRID HC**	**p-value for more edges than expected in HINT+HI2014**	**p-value for more edges than expected in iRefIndex14+KEGG**	**p-value for more edges than expected in ReactomeFI 2015**
2	Biliary-AdenoCA	0.5404	0.6994	0.6797	0.0117	0.6121	0.4985	0.4257	0.9883
3	Bladder-TCC	0.0156	0.0274	0.3042	0.7841	0.9883	0.9883	0.7504	0.2588
4	Bone-Leiomyo	0.0156	0.0000	0.0430	0.0118	0.9883	1.0000	0.9843	1.0000
5	Bone-Osteosarc	0.0000	0.0000	0.0039	0.0000	1.0000	1.0000	1.0000	1.0000
6	Breast-AdenoCa	0.1677	0.3344	0.9232	0.9547	0.8766	0.7395	0.0899	0.0490
7	Breast-LobularCa	0.6618	0.9452	0.9142	0.3370	0.5885	0.1760	0.1484	0.8075
8	CNS-GBM	0.0234	0.0000	0.1988	0.0000	0.9844	1.0000	0.8360	1.0000
9	CNS-Medullo	0.0273	0.3517	0.3696	0.8141	0.9961	0.8112	0.7516	0.2873
10	CNS-Oligo	0.8619	0.8906	1.0000	0.0431	0.3141	0.2978	0.0039	1.0000
11	ColoRect-AdenoCA	0.1158	0.1708	0.0078	0.0118	0.9148	0.8842	0.9922	0.9961
12	Eso-AdenoCa	0.5573	0.7943	0.8151	0.0853	0.5442	0.2795	0.2374	0.9264
13	Head-SCC	0.8905	0.9248	0.0312	0.4717	0.1705	0.1493	0.9922	0.6148
14	Kidney-RCC	0.8905	0.1489	0.1880	0.8167	0.2500	0.9315	0.8956	0.3359
15	Liver-HCC	0.0121	0.0000	0.1455	0.0384	0.9879	1.0000	0.8651	0.9620
16	Lung-AdenoCA	0.1054	0.0509	0.0800	0.0195	0.9298	0.9726	0.9396	0.9883
17	Lung-SCC	0.2367	0.1984	0.0702	0.4992	0.8416	0.8681	0.9454	0.5687
18	Lymph-BNHL	0.3056	0.5104	0.0196	1.0000	0.7491	0.5568	0.9804	0.0000
19	Ovary-AdenoCA	0.3910	0.3987	0.3015	0.6680	0.7028	0.6839	0.7389	0.4068
20	Panc-AdenoCA	0.4890	0.6224	0.0000	0.3632	0.5962	0.4429	1.0000	0.6726
21	Panc-Endocrine	0.0039	0.0118	0.0078	0.5899	0.9961	1.0000	1.0000	0.5901
22	Prost-AdenoCA	0.0431	0.3848	0.8032	0.0367	0.9726	0.7210	0.2617	0.9828
23	Skin-Melanoma	0.0960	0.5454	0.0069	0.4097	0.9388	0.5905	1.0000	0.6552
24	Stomach-AdenoCA	0.1754	0.0078	0.0436	0.6614	0.8733	1.0000	0.9731	0.3801
25	Uterus-AdenoCA	0.2223	0.2600	0.8185	0.5971	0.8455	0.7976	0.2245	0.4380

Summary of cohort results

For **each cohort**, are there **significantly more** or **signifcantly fewer interactions** between driver mutations than expected by chance?

- **9 cohorts** have significantly **fewer interactions** than expected ($p < 0.05$) in **multiple networks**:
  - Bladder-TCC, Bone-Leiomyo, Bone-Osteosarc, CNS-CBM, ColoRect-AdenoCA, Liver-HCC, Panc-Endocrine, Prost-AdenoCA, Stomach-AdenoCA

- **3 cohorts** have significantly **more interactions** than expected ($p < 0.05$) in **one network**:
  - Breast-AdenoCa, CNS-Oligo, Lymph-BNHL
  - Networks (iRefIndex14+KEGG and ReactomeFI 2015) are two denser.

Summary of patient results

For **each patient**, are there **significantly more or fewer interactions** between driver mutations than expected by chance?

- **3 patients** have significantly **fewer interactions** than expected ($p < 0.01$) in one network.
  - Lung-SCC patient ($p < 0.0001$ for fewer interactions):
    - Driver mutations in: CDKN2A, COL11A1, EGFR, FANCF, FAT1, FBXW7, IKZF2, KMT2D, NFE2L2, RHOA, SOX2, WWOX, ZFP36L1.
    - Induced subgraph of ReactomeFI 2015 has 0 interactions, vs. 5.98 interactions expected from null distribution.

- **19 patients** have significantly **more interactions** than expected ($p < 0.01$) in at least one network, where 3 of these patients have more interactions than expected in multiple networks.
  - Liver-HCC patient ($p < 0.0001$ for more interactions)
    - Driver mutations in CCND1, CDKN2A, PTEN, RB1, and TP53.
    - Subgraph of BioGrid HC has 4 interactions: CCND1-CDKN2A, CCND1-RB1, CDKN2A-TP53, PTEN-TP53 vs. 0.48 interactions expected from null distribution;.

Summary of interaction results

For **each pair of interacting genes**, do both genes contain driver mutations in **significantly more** or **significantly fewer patients** in a cohort than expected by chance?

- 16 interactions in BioGRID HC co-occur in significantly **fewer patients** within a cohort than expected by chance ($p < 0.05$):
  - ATM-TP53, ATRX-DAXX, **AXIN1-CTNNB1**, BCL2-TP53, BRCA2-TP53, CDK4-CDKN2A, CDK4-CDKN2B, CDK4-RB1, CDKN2A-MDM2, **CDKN2A-TP53**, CREBBP-MYC, ERG-SPOP, KDM6A-KMT2D, MAP3K1-TP53, MDM2-TP53, **PTEN-TP53**

- 19 interactions in BioGRID HC co-occur in significantly **more patients** within a cohort than expected by chance ($p < 0.05$):
  - **AKT1-CTNNB1**, AXIN1-MAP3K1, BRCA1-MYC, CCDC6-FBXW7, CCND1-CDK6, CCND1-CDKN2A, CCND2-RB1, **CDKN2A-TP53**, CUL1-FBXW7, ERBB2-PLCG1, ERBB4-GRB2, GPS2-NCOR1, MDM2-TERT, MYC-SMARCA4, NFKBIZ-STAT3, **PTEN-TP53**, RAF1-RB1, RB1-RBBP8, SMARCA4-TP53

Acknowledgements

David Haan (UCSC)
Jose MG Izarzugaza (DTU)
Abdullah Kahraman (UZH)
Kathleen Marchal (U. Ghent)
Sergio Pulido-Tamayo (U. Ghent)
Jüri Reimand (OICR)
Matthew Reyna (Princeton)
Cenk Sahinalp (U. Indiana, SFU)
Raunak Shrestha (UBC)
Miguel Vazquez (CNIO)
Lieven Verbeke (U. Ghent)

Members of PCAWG-5
Members of PCAWG-2-5-9-14