Computational analysis: Heterochromatin proteins and their link with genomic instability

Part 1: Link between oncogenic changes in heterochromatin proteins and the overall number of SVs

≥ 10 mutated / wildtype samples for it to be included

Deficiency or amplification, cancer type specific

SV calling by LINX

Overal the number of SVs increase upon alteration of heterochromatin proteins, mostly this is not cancer type specific or SVs type specific.

Correlations found:

|  |  |  |  |
| --- | --- | --- | --- |
| Gene | Mutation | Cancer | Remarks |
| ATRX | Deficiency | Bone  Soft tissue | *No difference in the amount of LINE elements?* |
| KDM6A | Deficiency | Lung  Urothelial  Any cancer types |  |
| DNMT3B | Amplification | Colorectal | Breakpoints at H3K9me2 |
| SETDB1 | Amplification | Lung  Urothelial |  |
| SIRT2 | Amplification | Urothelial |  |
| *LMNA* | *Amplification* | *Lung* | *Weak* |
| *SETDB1* | *Amplification* | *Breast* | *Weak* |

Comments:

* CBX5 and SUV39H1 did not pop up. Data that is available suggesting a link with heterochromatin instability is mostly biological data based on mutations/ deficiencies/ amplifications. No bio-informatics data.
* Often found correlations between heterochromatin like chromatin and the amount of mutations/ SNVs. H3K9me3 etc. rich areas have an increased amount of mutations. Is this correlations also visible in the HMF dataset?
* LINX annotations: <https://github.com/hartwigmedical/hmftools/tree/master/sv-linx>

Next steps:

* Look at whole chromosome and segmental aneuploidies (chromosome arm gains/losses). Whole arm gains or losses are specifically interesting or each, possibly indicating problems in the pericentromeric heterochromatin.

Part 2: Chromosome arm gain/losses in samples with deficiencies/amplifications of heterochromatin proteins

Input number of chromosome arm gains and losses and which arms have copy number gain/loss compared to genome copy number. Required ≥ 5 mutated / wildtype samples for it to be included.

Deficiencies: biallelic loss

Focal amplifications: gene copy number ≥ genome copy number

Within a sample:

* A gene deficient is only considered deficient if the gene has no amplification
* Similarly a gene is only considered amplified if the gene has no deficiencies
* A gene must have no deficiency or amplification to be considered WT

Samples that do not meet the above criteria will fall into the ‘grey’ list (excluded from analysis for the target gene).

Chromosome arm and genome copy number were defined in two different features, n\_arm.gain/loss, which is the number of arms that have an increase/decrease in copy number compared to the genome copy number, and gain.3q or similarily named features which is a logical feature simply stating for that indicated arm whether the arm has an increase/decrease in copy number compared to the genome copy number. Both features just state if there is an increase/decrease compared to genome copy numbers but not the extend of the change, so there is no distinction between copy number 3 and 5, both are considered as just an increase.

Enrichment analysis is a list with each row containing one statistical comparison (mut vs. wildtype). The total number of rows comes from doing the statistical comparisons for all conditions: number of cancer types x number of genes x def/amp x number of features.

cid:image002.png@01D7223A.95D041A0

So in the above line the following question is asked: Within the ovarian cancer samples, do samples with amplified DNMT1 have significantly more arm gains **overall** compared to samples with WT DNMT1? Here the features are numeric, so I test this with a Wilcox test (basically a t-test, but when you don’t know if your data follows a bell curve). For the gain.3q, etc features, those are of logical type, so I test significance with a Fisher’s exact test.

Found focal amplifications of many genes, SIRT2 and SETDB1 included, leading to more arm gains.

Part 3: Amount of LINE elements correlated with deficiencies/amplifications of heterochromatin proteins

Correlations found:

|  |  |  |  |
| --- | --- | --- | --- |
| Gene | Mutation |  | Remarks |
| ATRX | Deficiency | Chromosome arm aneuploidy (both losses and gains) | *weakly* |
| ATRX | Deficiency | LINE high samples | *Weakly enriched (possibly specific for lung)* |
| SIRT2 | Amplification | Chromsome arm gains | *Potentially* |
| SIRT2 | Amplification | LINE high samples | *Enriched (possibly specific for lung)* |
| SETDB1 | Amplification | Chromsome arm gains | *Potentially* |
| SETDB1 | Amplification | LINE high samples | *Enriched (possibly specific for lung)* |
| KDM6A | Deficiency | High LINE elements enrichment | *Likely specific for lung* |

Genes appear in both enrichment analysis (2 lines of evidence) as well as appearing in multiple cancer types (so no cancer specific effects).

Luan is not overly convinced, says it’s hard to delink cancer types and the specific aneuploidies seen. There is a possibility to do instead of a clear cutoff of 10 LINE elements, determine a cutoff (e.g. top 10%) per cancer type.

Part 4: Ranking