Pathway analysis of lung cancer associated variants

Xavier Duran, Iván Galván
GCAT Genomes for Life
Institut de Recerca Germans Trias i Pujol (IGTP)

HealthForecast
December 13th 2016

Gene Set Analysis (GSA)

Overrepresentation analysis

Given a list of genes, are any of the pathways (Gene Ontologies) *surprisingly* enriched in that list?

How? Calculating p-value with a simple hypergeometric distribution

	Significant genes	Non-significant genes	Total
genes in the group	k	K-k	K
other genes	n-k	N-K-n+k	N-K
total	n	N-n	N

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

From variants to genes

Variant list

name	chrom	chromStart	chromEnd	stability	scaled_coefficients	ref	alt
12:65995405:C:G	12	65995405	65995406	10	0.6901523	C	G
12:132695710:G:A	12	132695710	132695711	10	0.6184025	G	A
12:122307580:A:G	12	122307580	122307581	6	0.5574568	A	G
12:86433258:A:C	12	86433258	86433259	8	0.5441592	C	A
12:63058957:A:G	12	63058957	63058958	10	0.5285514	G	A
12:121149596:A:G	12	121149596	121149597	6	0.5188642	G	A
12:127933561:C:CT	12	127933561	127933562	10	0.5086084	C	CT
12:26860830:CA:C	12	26860830	26860831	9	0.4982072	C	CA
12:64966695:T:C	12	64966695	64966696	10	0.4958828	C	T
12:78538314:C:T	12	78538314	78538315	8	0.4950422	T	C
12:127933756:A:G	12	127933756	127933757	10	0.4940085	A	G
12:118388554:A:T	12	118388554	118388555	8	0.4931533	T	A
12:2011473:C:A	12	2011473	2011474	4	0.4925469	C	A
12:614204:A:G	12	614204	614205	8	0.4921972	G	A
12:52954624:C:A	12	52954624	52954625	6	0.4888366	A	C
12:64963697:G:A	12	64963697	64963698	9	0.4887630	A	G
12:91262911:T:C	12	91262911	91262912	8	0.4886846	C	T
12:117985667:C:T	12	117985667	117985668	8	0.4884988	T	C
12:18075974:G:A	12	18075974	18075975	3	0.4883262	A	G
12:125500901:G:GT	12	125500901	125500902	8	0.4879617	GT	G

Table 2: Variant list

From variants to genes

One of several well known browsers for the retrieval of annotated genomic information.

Assembly GRCh37 (hg19)

- cut chromosome into pieces
- sequence those pieces
- put them together

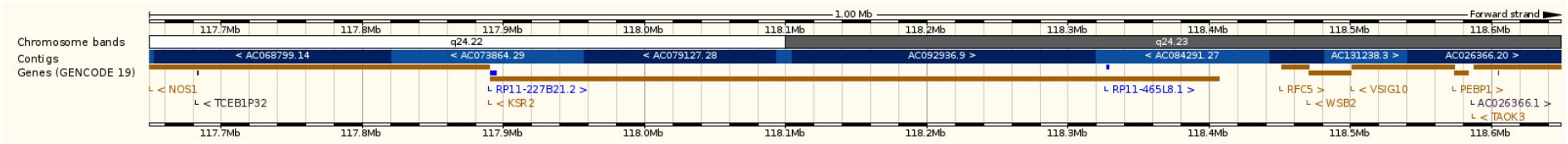
From variants to genes

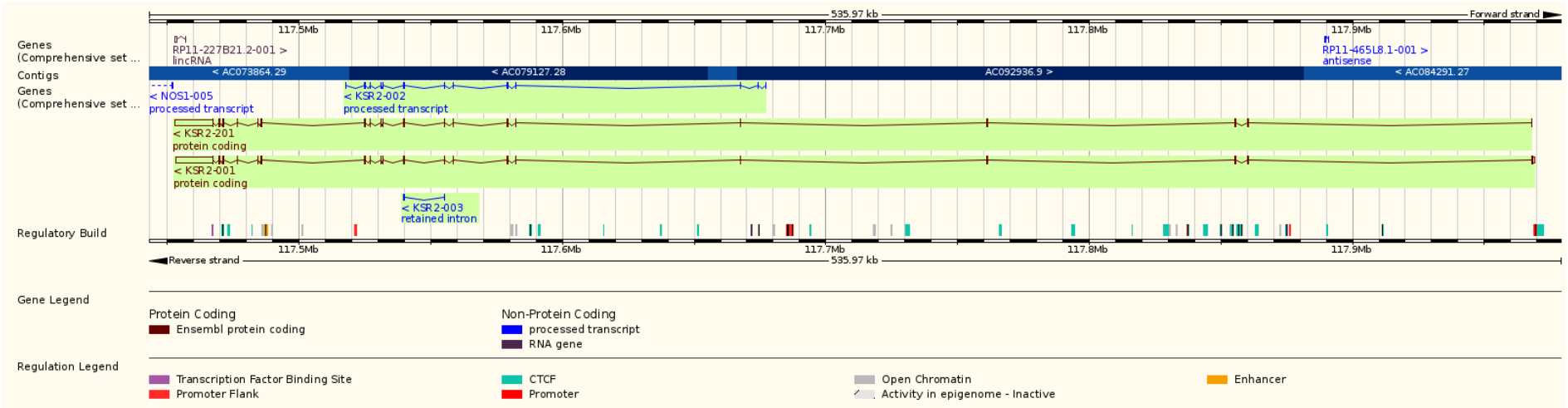
Chromosome 12

Length (bps)	133,851,895
Coding genes	1,071
Non coding gene count	1,131
Pseudogenes	616
Short Variants	6,980,191

From variants to genes

Region of gene KSR2





From variants to genes

Seq2pathway

Seq2pathway is an R/Python wrapper for pathway (or functional gene-set) analysis of genomic loci, adapted for advances in genome research.

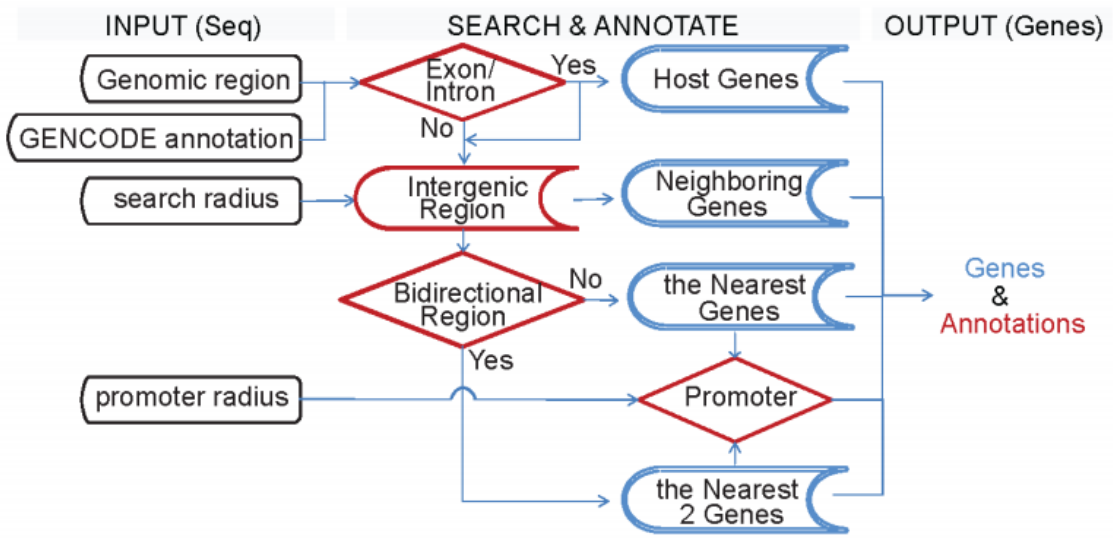


Figure 1: seq2gene workflow

From variants to genes

Code

```
gene_dat <- runseq2gene(
  inputfile = snp_dat,
  genome = "hg19",
  adjacent = TRUE,
  SNP = TRUE,
  search_radius = 150000,
  PromoterStop = FALSE,
  NearestTwoDirection = TRUE
)
```

From variants to genes

Variants and nearby genes

name	chrom	chromStart	chromEnd	type	gene_name	source
12:116867666:C:T	12	116867666	116867667	Nearest_R	MAP1LC3B2	protein_coding
12:76301407:G:C	12	76301407	76301408	Intron	RP11-114H23.1	lincRNA
12:53079192:C:A	12	53079192	53079193	Nearest_R	KRT77	protein_coding
rs4325389	12	10176700	10176701	Nearest_R	CLEC9A	protein_coding
12:20296941:C:G	12	20296941	20296942	Nearest_L	AEBP2	protein_coding
12:26143130:T:G	12	26143130	26143131	Intron	RASSF8	protein_coding
12:17770479:C:T	12	17770479	17770480	Nearest_L	LMO3	protein_coding
12:92733101:T:C	12	92733101	92733102	Nearest_L	BTG1	protein_coding
rs10847773	12	129597700	129597701	Exon	RP11-669N7.2	antisense
12:67353131:T:A	12	67353131	67353132	Nearest_L	GRIP1	protein_coding
12:20064967:G:A	12	20064967	20064968	Intron	RP11-405A12.2	lincRNA
12:17605648:G:C	12	17605648	17605649	Nearest_R	PIK3C2G	protein_coding
12:101132821:T:TTA	12	101132821	101132822	Intron	ANO4	protein_coding
12:84363365:ATTT:ATT	12	84363365	84363366	Nearest_L	TMTC2	protein_coding
rs10772825	12	15101478	15101479	Intron	ARHGDIB	protein_coding
12:131280130:T:A	12	131280130	131280131	Intron	STX2	protein_coding
12:68102814:T:C	12	68102814	68102815	Nearest_R	IFNG	protein_coding
rs7306456	12	132703218	132703219	Intron	GALNT9	protein_coding
12:26188369:TATATA:T	12	26188369	26188370	Intron	RASSF8	protein_coding
12:95106431:C:CCA	12	95106431	95106432	Nearest_L	TMCC3	protein_coding

Table 4: Variant annotation list

From variants to genes

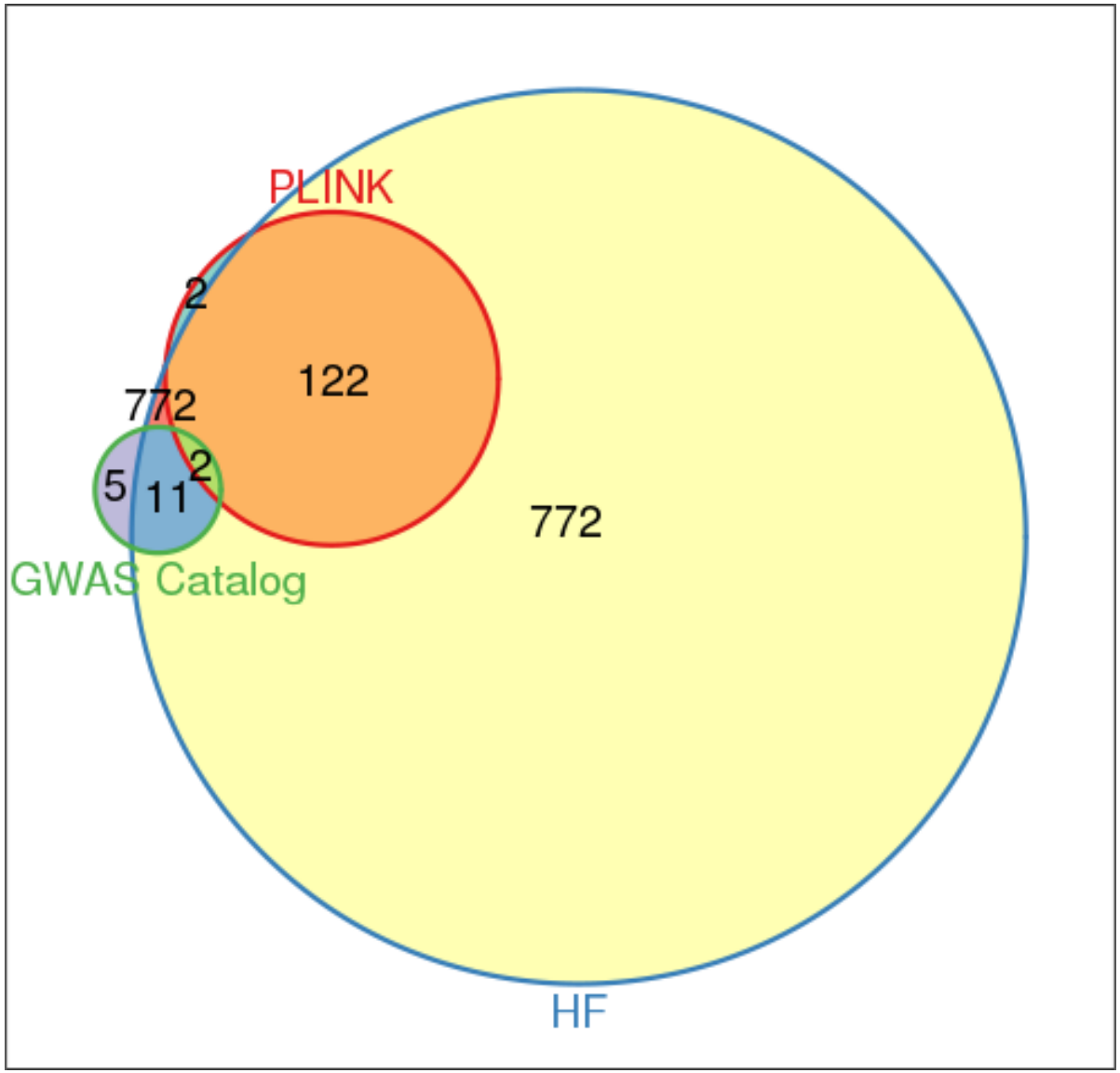
Gene list

gene	frequency
CCDC91	360
TMEM132B	345
TMEM132C	340
ATF7IP	333
IFLTD1	326
ANKS1B	303
GRIP1	265
AEBP2	247
PPFIA2	247
NEDD1	229
RP11-181C3.1	214
SOX5	188
NAV3	161
AC092850.1	160
C12orf79	157
TMEM132D	156
CMKLR1	155
DIP2B	148
SFSWAP	148
CRADD	141

Table 5: Gene frequency list

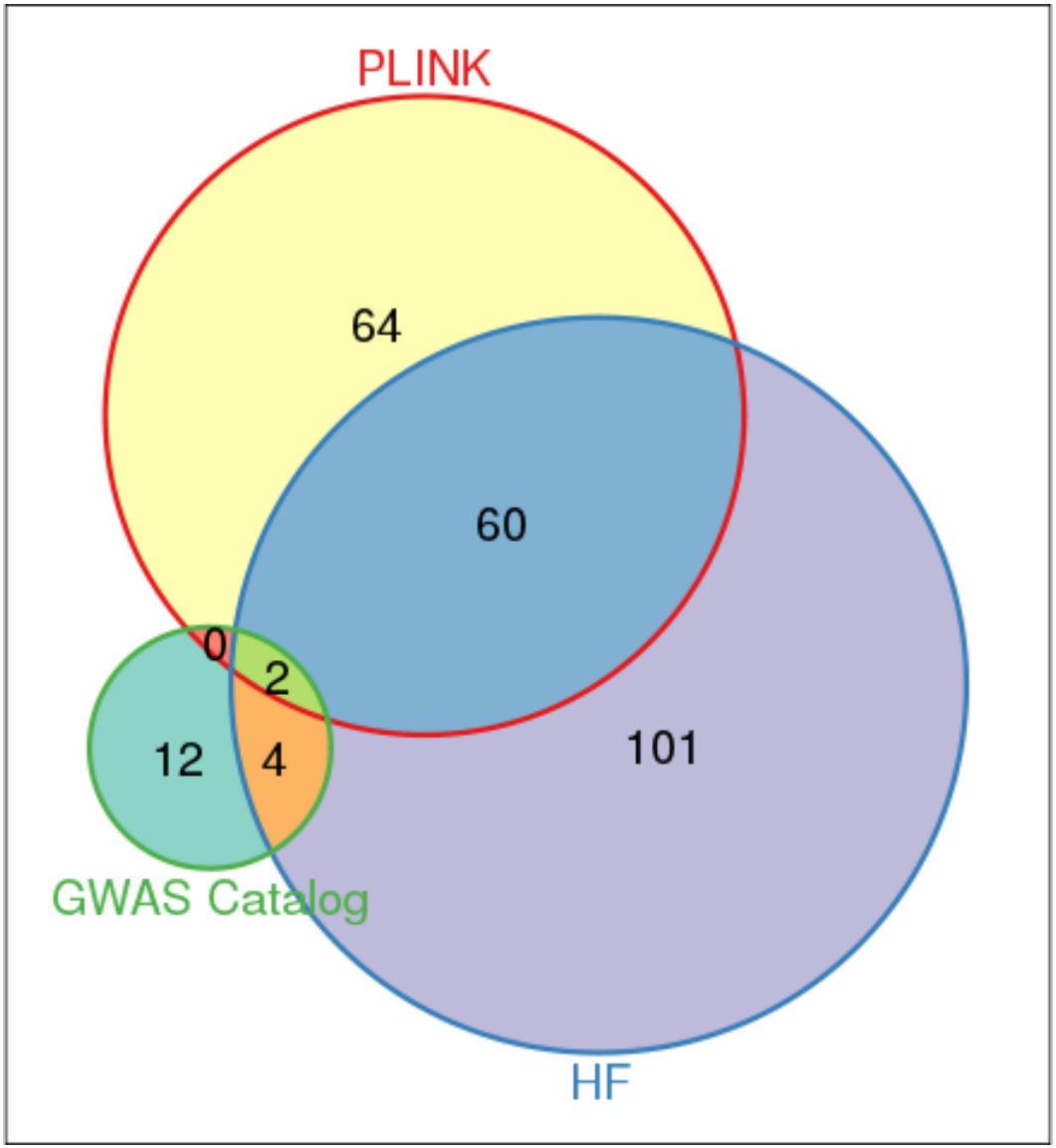
From variants to genes

Sets intersection



From variants to genes

Sets intersection



From genes to pathways

Reactome

Reactome is a free, open-source, curated and peer reviewed pathway database

UniProt	proteins
Ch EBI	small molecules
Ensembl	genes and transcripts
RNA Central	ncRNAs
PubMed	literature evidence
Gene Ontology (GO)	molecular function, biological process

From genes to pathways

Health Forecast selected variants

Pathway.identifier	Pathway.name	Entities.pValue	Submitted.entities.found
R-HSA-6809371	Formation of the cornified envelope	8.664749734155208E-6	KRT71;KRT82;KRT80;KRT4;KRT3;KRT2;KRT1;KRT79;KRT8;KRT78;KRT7;KRT77;KRT76;KRT75;KRT86;KR
R-HSA-420499	Class C/3 (Metabotropic glutamate/pheromone receptors)	0.004953950850458244	TAS2R7;TAS2R42
R-HSA-6805567	Keratinization	0.014321908213508316	KRT71;KRT82;KRT80;KRT4;KRT3;KRT2;KRT1;KRT79;KRT8;KRT78;KRT7;KRT77;KRT76;KRT75;KRT86;KR
R-HSA-72731	Recycling of eIF2:GDP	0.03954795734154837	SLC5A8;EIF2B1;EIF2S3L
R-HSA-3249367	STAT6-mediated induction of chemokines	0.054077331642961024	TBK1;STAT6
R-HSA-6802948	Signaling by high-kinase activity BRAF mutants	0.05887835859772694	RAP1B;VWF;PEBP1;KRAS;KSR2
R-HSA-5625900	RHO GTPases activate CIT	0.06381045902536409	CDKN1B;MYL6B;CIT
R-HSA-6802946	Signaling by moderate kinase activity BRAF mutants	0.06441268586272786	RAP1B;VWF;PEBP1;KRAS;KSR2
R-HSA-5674135	MAP2K and MAPK activation	0.06441268586272786	RAP1B;VWF;PEBP1;KRAS;KSR2
R-HSA-6802949	Signaling by RAS mutants	0.0660408405716818	RAP1B;VWF;PEBP1;RASAL1;KRAS;KSR2
R-HSA-6802955	Paradoxical activation of RAF signaling by kinase inactive BRAF	0.07025180102062634	RAP1B;VWF;PEBP1;KRAS;KSR2
R-HSA-139853	Elevation of cytosolic Ca2+ levels	0.08356236576501674	P2RX7;P2RX4;P2RX2;ITPR2
R-HSA-5250989	Toxicity of botulinum toxin type G (BoNT/G)	0.09637150792363747	SYT1;VAMP1
R-HSA-69091	Polymerase switching	0.1180671078261919	RFC5;PRIM1
R-HSA-69109	Leading Strand Synthesis	0.1180671078261919	RFC5;PRIM1
R-HSA-2995410	Nuclear Envelope Reassembly	0.1180671078261919	ANKLE2;LEMD3;TMPO
R-HSA-2995383	Initiation of Nuclear Envelope Reformation	0.1180671078261919	ANKLE2;LEMD3;TMPO
R-HSA-5635851	GLI proteins bind promoters of Hh responsive genes to promote transcription	0.12008284192932228	GLI1
R-HSA-8849470	PTK6 Regulates Cell Cycle	0.12008284192932228	CDKN1B;CDK4
R-HSA-110328	Recognition and association of DNA glycosylase with site containing an affected pyrimidine	0.12008284192932228	SMUG1;TDG

Table 7: Pathway analysis with HF variants

From genes to pathways

PLINK & HF

Pathway.identifier	Pathway.name	Entities.pValue	Submitted.entities.found
R-HSA-6802948	Signaling by high-kinase activity BRAF mutants	0.002558965840406824	VWF;KRAS;KSR2
R-HSA-6802946	Signaling by moderate kinase activity BRAF mutants	0.0027629322765093667	VWF;KRAS;KSR2
R-HSA-5674135	MAP2K and MAPK activation	0.0027629322765093667	VWF;KRAS;KSR2
R-HSA-6802955	Paradoxical activation of RAF signaling by kinase inactive BRAF	0.002977768630790245	VWF;KRAS;KSR2
R-HSA-6802949	Signaling by RAS mutants	0.005445282761517123	VWF;KRAS;KSR2
R-HSA-3656243	Defective ST3GAL3 causes MCT12 and EIEE15	0.005553893432453161	LUM;KERA
R-HSA-3656244	Defective B4GALT1 causes B4GALT1-CDG (CDG-2d)	0.005553893432453161	LUM;KERA
R-HSA-3656225	Defective CHST6 causes MCDC1	0.005553893432453161	LUM;KERA
R-HSA-6802952	Signaling by BRAF and RAF fusions	0.009959479808282934	VWF;KRAS;KSR2
R-HSA-2428933	SHC-related events triggered by IGF1R	0.014711331693763041	KRAS;IGF1

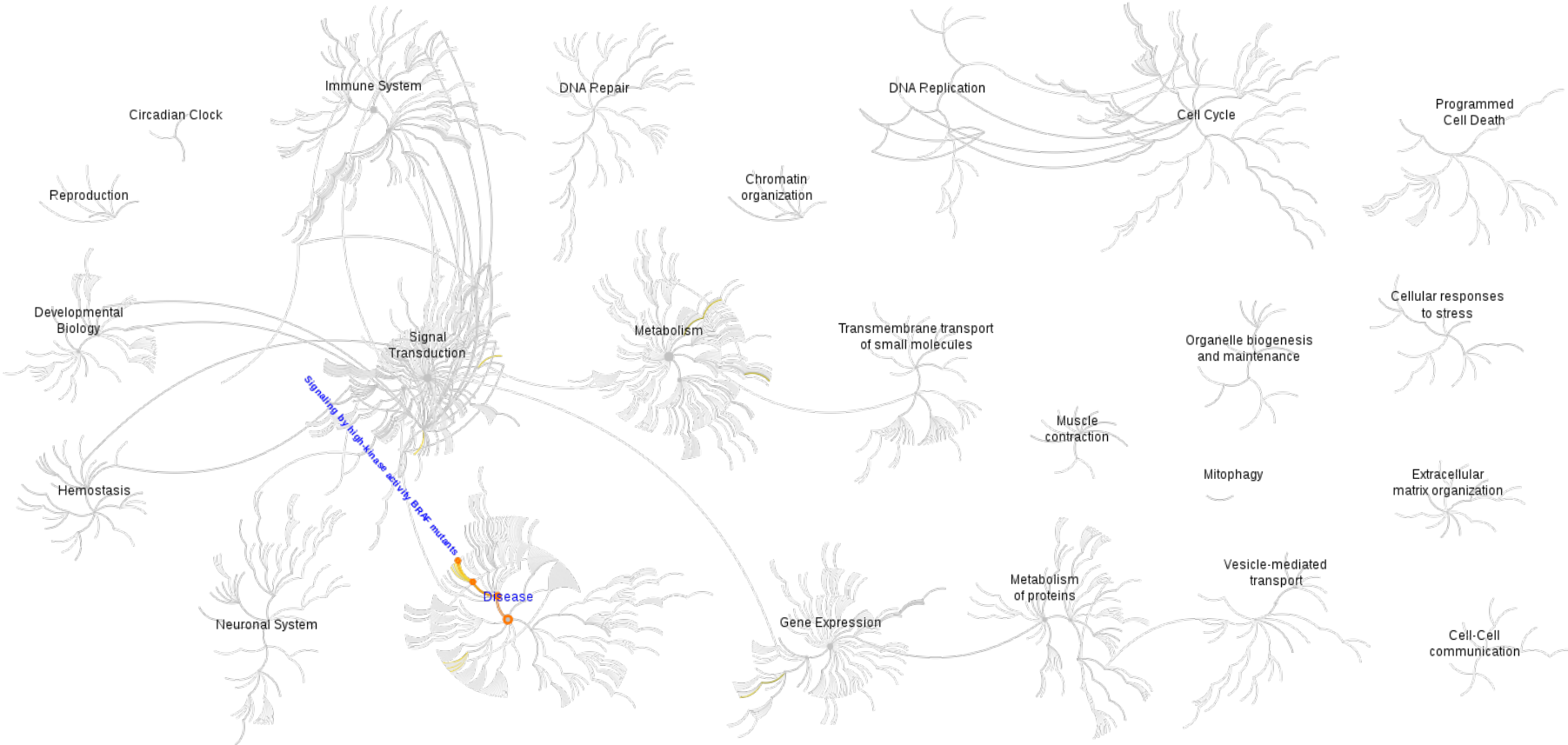
Table 8: Pathway analysis with HF variants, stability > 5

Pathway.identifier	Pathway.name	Entities.pValue	Submitted.entities.found
R-HSA-6802948	Signaling by high-kinase activity BRAF mutants	0.008027391862395206	VWF;KSR2
R-HSA-6802946	Signaling by moderate kinase activity BRAF mutants	0.008508534510688448	VWF;KSR2
R-HSA-5674135	MAP2K and MAPK activation	0.008508534510688448	VWF;KSR2
R-HSA-6802955	Paradoxical activation of RAF signaling by kinase inactive BRAF	0.009006296531779867	VWF;KSR2
R-HSA-6804116	TP53 Regulates Transcription of Genes Involved in G1 Cell Cycle Arrest	0.012423880473841464	E2F7
R-HSA-428790	Facilitative Na+-independent glucose transporters	0.012423880473841464	SLC2A3
R-HSA-6802949	Signaling by RAS mutants	0.014251144847456931	VWF;KSR2
R-HSA-6802952	Signaling by BRAF and RAF fusions	0.022584674256351356	VWF;KSR2
R-HSA-6802957	Oncogenic MAPK signaling	0.038532765593673135	VWF;KSR2
R-HSA-3249367	STAT6-mediated induction of chemokines	0.04305190862841446	TBK1

Table 9: Pathway analysis with PLINK variants

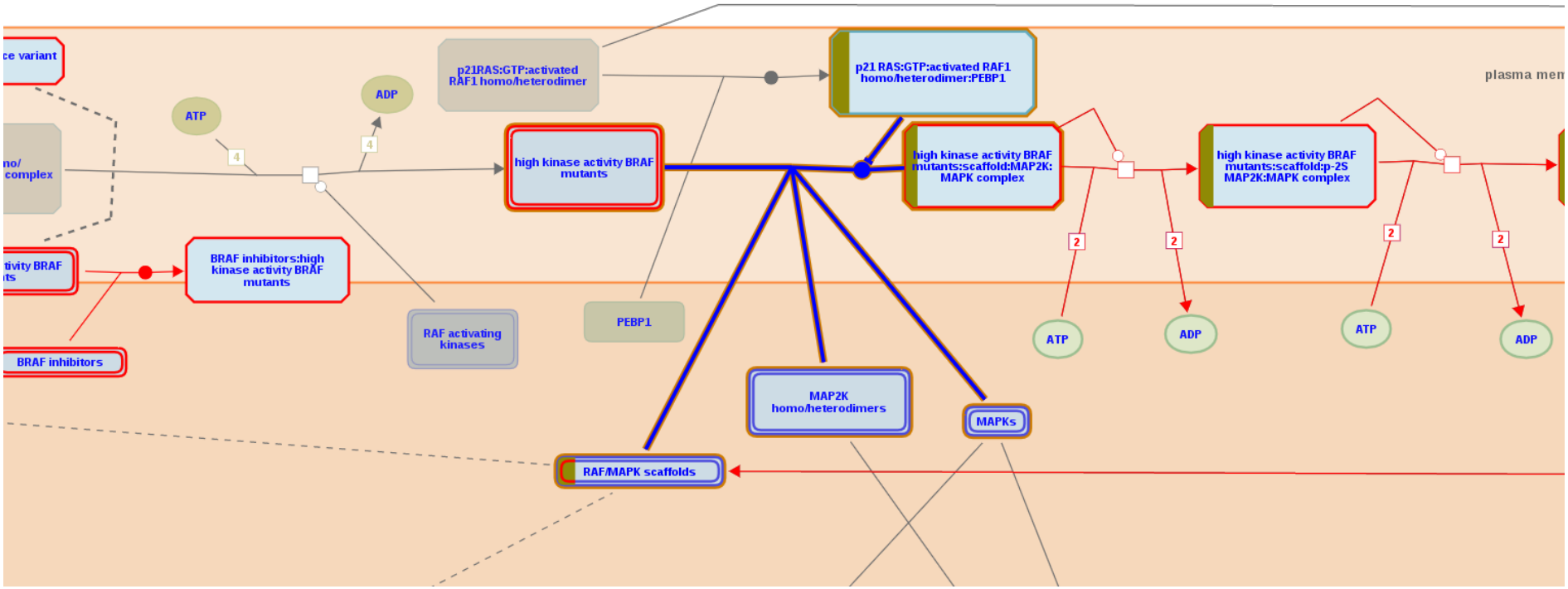
From genes to pathways

Signaling by high-kinase activity BRAF mutants



From genes to pathways

Signaling by high-kinase activity BRAF mutants



From genes to pathways

Curve ROC analysis for the top pathway

	name	gene_name	p-value
1	12:25371462:C:T	KRAS	0.033332600
2	12:25409506:G:A	KRAS	0.002104760
3	12:25528516:A:AT	KRAS	0.001517350
4	12:25560773:GA:G	KRAS	0.201970000
5	12:118309635:A:AT	KSR2	0.028058900
6	12:118387112:G:GA	KSR2	0.078734500
7	12:118396478:C:T	KSR2	0.015582300
8	rs7972611	KSR2	0.056434400
9	rs7977174	KSR2	0.000350504
10	12:6087041:C:A	VWF	0.003123620
11	12:6272372:T:C	VWF	0.060201300

Table 10: Selected SNPs for the ROC curve analysis after LD filter applied

From genes to pathways

Curve ROC analysis for the top pathway

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.1111	1.5217	-2.70	0.0069
ECOG1	-0.4369	0.6291	-0.69	0.4874
ECOG2	-15.3020	4406.4471	-0.00	0.9972
fumador2	-15.2612	1823.9592	-0.01	0.9933
histologia2	1.7499	0.6621	2.64	0.0082
histologia3	-16.2795	2307.7481	-0.01	0.9944
histologia4	2.2762	1.0550	2.16	0.0310
tractament2	-1.3213	0.7536	-1.75	0.0796
tractament3	-1.9501	1.3699	-1.42	0.1546
sex2	-0.4411	0.7996	-0.55	0.5812
12:6087041:C:A	1.2234	0.4368	2.80	0.0051
12:6272372:T:C	0.5293	0.3847	1.38	0.1689
12:25371462:C:T	0.8230	0.4414	1.86	0.0622
12:25409506:G:A	-1.1204	0.5417	-2.07	0.0386
12:25528516:A:AT	0.9936	0.4680	2.12	0.0337
12:25560773:GA:G	-0.2570	0.4595	-0.56	0.5760
rs7977174	1.2235	0.4492	2.72	0.0065
rs7972611	-0.4810	0.4364	-1.10	0.2704
12:118309635:A:AT	-1.1496	0.4273	-2.69	0.0071
12:118387112:G:GA	-0.6293	0.4730	-1.33	0.1834
12:118396478:C:T	1.2912	0.4460	2.90	0.0038

Table 11: Model 1: Logistic model with clinical variables + SNPs

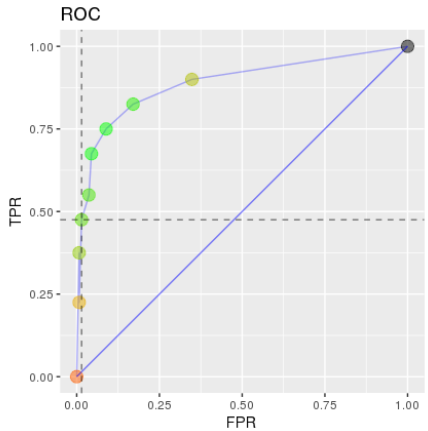


Figure 2: ROC curve model 1 (AUC = 0.92)

From genes to pathways

Curve ROC analysis for the top pathway

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1923	0.3985	-2.99	0.0028
sex2	-0.5842	0.5731	-1.02	0.3080
ECOG1	-0.2326	0.4097	-0.57	0.5702
ECOG2	-16.3738	2797.4420	-0.01	0.9953
fumador2	-15.9061	1224.6130	-0.01	0.9896
tractament2	-0.4486	0.4716	-0.95	0.3414
tractament3	0.0126	0.7348	0.02	0.9863
histologia2	0.7955	0.4094	1.94	0.0520
histologia3	-16.0734	1599.1012	-0.01	0.9920
histologia4	1.8973	0.7537	2.52	0.0118

Table 12: Model 2: Logistic model with clinical variables

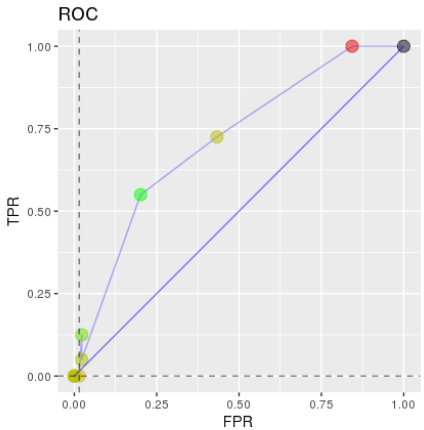


Figure 3: ROC curve model 2 (AUC=0.71)

Summary

The functional analysis is incomplete and unreliable because only chromosome 12 has been analyzed

More restrictive signification threshold led to known associated biological processes to lung cancer

Less restrictive ones let other biological processes (like keratinization) come to the surface