

Consensus Machine-Learning Signature for Post-DAA HCV-Associated Hepatocellular Carcinoma Risk from Liver Transcriptomic Profiles

Dilnaz Abdirakhimova

Capstone Project, Spring 2026

Supervisor: Dongming Wei

Co-Supervisor: Syed Hani Hassan Abidi

April 20, 2026

Abstract

Background. Many patients with hepatitis C virus (HCV)-related cirrhosis remain at risk of hepatocellular carcinoma (HCC) after antiviral therapy. Transcriptomic markers that stratify future HCC risk from liver tissue could support risk-adapted surveillance, complementing standard imaging and serum tests.

Methods. We re-analysed the public GSE15654 dataset [Hoshida et al., 2013]: genome-wide expression from liver biopsies of 216 patients with HCV cirrhosis, followed prospectively for a median of ten years. Twenty-one patients with fewer than five years of follow-up were excluded, leaving 195 individuals (65 HCC, 130 no HCC), partitioned into discovery ($n = 136$) and validation ($n = 59$) strata (70/30, stratified). We trained three classifiers (Random Forest, elastic-net logistic regression, linear SVM). Univariate gene pre-selection was performed inside each cross-validation fold to limit leakage between training and test folds. A consensus gene set was defined by majority overlap of the top 30 genes per model.

Results. On the discovery cohort, stratified 5-fold cross-validation yielded mean AUROCs of 0.60 (± 0.05) for Random Forest, 0.55 (± 0.06) for SVM, and 0.50 (± 0.06) for elastic-net. On the held-out validation cohort, AUROCs were 0.62 (95% CI: 0.46–0.79), 0.61 (0.44–0.76), and 0.60 (0.44–0.76) for Random Forest, SVM, and a simple average of the three models' predicted probabilities, respectively. Sixteen genes met the consensus rule; *AZU1* and *PHF17* appeared in all three top-30 lists.

Conclusions. Performance was modest but directionally consistent across models and splits, which aligns with predicting a rare future event from a single pre-

malignant biopsy under high dimensionality. The 16-gene panel highlights immune- and chromatin-related biology as candidates for replication in larger, ideally multi-centre, post-therapy cohorts.

Keywords: hepatocellular carcinoma, hepatitis C virus, cirrhosis, HCC risk prediction, machine learning, consensus biomarker, gene expression, transcriptomics

Contents

1	Introduction	4
1.1	Objectives	4
2	Literature Review	5
2.1	HCC Risk in the Post-DAA Era	5
2.2	Transcriptomic Signatures for HCC Risk Stratification	5
2.3	Consensus Machine Learning in Genomics	5
3	Data and Methods	5
3.1	Dataset: GSE15654	5
3.2	Data Preprocessing	6
3.3	Univariate Gene Pre-Selection (Inside CV)	6
3.4	Machine Learning Models	7
3.5	Consensus Biomarker Panel (Objective 2)	7
3.6	Validation and Evaluation (Objective 3)	7
4	Results	8
4.1	Discovery Cohort: Cross-Validated Performance (Objective 1)	8
4.2	Gene-Set Stability and Feature Importance (Objective 1)	9
4.3	Consensus Biomarker Panel (Objective 2)	11
4.4	Independent Validation (Objective 3)	13
5	Discussion	15
5.1	Performance in Context	15
5.2	Biological Interpretation of Consensus Genes	16
5.3	Comparison with Hoshida et al. (2013)	16
5.4	Model-Specific Behaviour	16
6	Limitations	17
7	Future Work	17
8	Conclusion	18
A	Code Excerpts	19
A.1	Data Loader (Excerpt)	19
A.2	Nested CV with Univariate Filter (Excerpt)	20
A.3	Consensus Derivation (Excerpt)	20

1 Introduction

Hepatocellular carcinoma (HCC) ranks among the most common malignancies globally and a leading cause of cancer mortality [Sung et al., 2021]. Chronic hepatitis C virus (HCV) infection is an important risk factor: HCV-related cirrhosis carries roughly a 1–8% annual incidence of HCC [El-Serag, 2012]. Direct-acting antiviral (DAA) therapy now achieves sustained virological response (SVR) in more than 95% of treated individuals [Falade-Nwulia et al., 2017]. Cirrhosis nonetheless confers residual risk after SVR, on the order of 1–3% per year in many reports, so surveillance remains relevant [Ioannou et al., 2020, Reig et al., 2016].

Ultrasound, with or without serum alpha-fetoprotein (AFP), remains the backbone of surveillance but has limited sensitivity in meta-analyses (often roughly 47–63%) [Tzartzeva et al., 2018]. Molecular markers that refine risk among patients with cirrhosis could help target intensified follow-up to those who benefit most, while avoiding unnecessary testing where yield is low.

Genome-wide profiling of cirrhotic liver captures heterogeneity that may precede malignant transformation. Hoshida et al. showed that a 186-gene signature originally derived from HCC tissue could stratify cirrhosis patients by subsequent outcome [Hoshida et al., 2013]. Which genes recur across independent statistical learners is still open, because each method stresses different aspects of high-dimensional expression data.

We therefore compared three standard classifiers: Random Forest (non-linear, ensemble-based importance), elastic-net logistic regression (sparse linear coefficients), and a linear support vector machine (margin-based weights). Intersecting gene rankings yields a short consensus list that is less tied to the idiosyncrasies of any one algorithm.

1.1 Objectives

This capstone addresses three aims:

- Objective 1: Train and compare three ML classifiers** (Random Forest, Elastic-Net logistic regression, linear SVM) on legacy HCV cirrhosis transcriptomic profiles to predict HCC development, evaluating gene-set stability and predictive performance.
- Objective 2: Derive a consensus biomarker panel** (≤ 20 genes) by majority-vote intersection of the top gene lists from each model.
- Objective 3: Validate the consensus signature** on a held-out cohort and assess reproducibility via bootstrap confidence intervals.

2 Literature Review

2.1 HCC Risk in the Post-DAA Era

Widespread DAA use has changed HCV care, yet HCC remains a concern after SVR in patients with advanced fibrosis or cirrhosis. Population studies suggest substantial relative risk reduction after cure, while absolute risk in cirrhotic patients can remain appreciable [Ioannou et al., 2020]. Residual risk plausibly reflects fibrosis, immune remodelling, and clonal evolution that are not fully reversed by viral clearance [Hoshida, 2014].

2.2 Transcriptomic Signatures for HCC Risk Stratification

Hoshida et al. (2013) showed that expression in cirrhotic liver biopsy relates to later clinical events, including HCC, using long follow-up in an HCV cirrhosis cohort [Hoshida et al., 2013]. Later work linked immune infiltration and epigenetic change in cirrhosis to progression risk [Kim et al., 2017, Nakagawa et al., 2016]. Because each study typically relied on one modelling strategy, cross-method agreement on gene lists is still limited.

2.3 Consensus Machine Learning in Genomics

Combining outputs from several learners is a common way to reduce dependence on a single inductive bias [Chen et al., 2014]. Genes that rank highly under Random Forest importance, elastic-net shrinkage, and SVM margins are arguably more likely to reflect reproducible signal than genes favoured by only one pipeline, though overlap will shrink as noise dominates.

Random Forests summarise contribution via impurity-based importance; elastic-net estimation yields sparse coefficients; linear SVMs emphasise margin-normalised weights. Their intersection defines a compact panel for biological follow-up.

3 Data and Methods

3.1 Dataset: GSE15654

We used the publicly available GSE15654 dataset [Hoshida et al., 2013], deposited in the NCBI Gene Expression Omnibus (GEO). This dataset comprises genome-wide expression profiles from liver biopsies of 216 patients with HCV-related early-stage (Child-Pugh class A) cirrhosis, prospectively followed at an Italian clinical centre for a median of 10 years.

- **Platform:** Illumina HumanRef-8 v3.0 (GPL8432), measuring 24,526 probes.
- **Outcome:** Binary HCC development during follow-up (1 = developed HCC, 0 = no HCC).

- **Sample size after quality control:** Of 216 patients, 21 no-HCC patients with less than 5 years of follow-up were excluded to reduce label noise from censoring, yielding 195 patients (65 developed HCC, 130 did not).

3.2 Data Preprocessing

Preprocessing was performed in Python using NumPy, Pandas, and Scikit-learn:

1. **Probe-to-gene mapping:** Probes were mapped to gene symbols using the GPL8432 platform annotation. Multi-probe genes were collapsed by averaging expression values, yielding 18,402 unique genes.
2. **Missing value handling:** Samples with >30% missing values were excluded; remaining missing values were imputed with per-gene medians.
3. **Low-expression filter:** Genes with mean expression below 0.5 ($\log_2$ scale) were removed.
4. **Low-variance filter:** Genes in the bottom 20th percentile of variance were removed, retaining 14,721 genes.
5. **Cohort split:** A stratified 70/30 train-test split (preserving HCC prevalence) created a discovery cohort ($n = 136$; 45 HCC) and a validation cohort ($n = 59$; 20 HCC).

3.3 Univariate Gene Pre-Selection (Inside CV)

Given the extreme dimensionality ($p = 14,721$ genes, $n = 136$ samples), a univariate pre-selection step was applied *inside each cross-validation fold* to prevent data leakage:

1. For each CV training fold, a two-sided Mann-Whitney U test was computed for every gene comparing HCC-positive vs. HCC-negative patients.
2. The top 500 genes (ranked by ascending p -value) were retained for that fold.
3. Standardisation (z -score) was fitted on the training fold and applied to the test fold.

This nested workflow limits optimism bias from filtering on the full discovery set before cross-validation; reported CV AUROCs should therefore more closely track held-out behaviour than those from a filter fit once outside CV.

3.4 Machine Learning Models

Three complementary classifiers were trained:

1. **Random Forest (RF):** 1,000 trees, max depth 5, min 5 samples per leaf, square-root feature subsampling, balanced class weights.
2. **Elastic-Net Logistic Regression:** $C = 0.5$, ℓ_1 -ratio = 0.3 (SAGA solver), balanced class weights, 10,000 max iterations.
3. **Linear SVM:** $C = 0.1$, balanced class weights, 10,000 max iterations; probabilities obtained via sigmoid calibration.

All models used stratified 5-fold cross-validation on the discovery cohort. Feature importance was extracted as:

- RF: Mean decrease in Gini impurity, averaged across folds.
- Elastic-Net: Absolute coefficient magnitude, averaged across folds.
- SVM: Absolute weight vector magnitude, averaged across folds.

After CV, each model was refitted on the full discovery cohort (with univariate filter on full discovery data + standardisation) for validation.

3.5 Consensus Biomarker Panel (Objective 2)

The top-30 genes from each model were identified. A gene was included in the consensus panel if selected by ≥ 2 of 3 models (majority vote). Genes were ranked by (1) number of votes (descending) and (2) mean importance rank across models (ascending). The panel was capped at 20 genes.

3.6 Validation and Evaluation (Objective 3)

Each model was evaluated on the held-out validation cohort ($n = 59$). Performance metrics included:

- Area Under the Receiver Operating Characteristic Curve (AUROC)
- 95% bootstrap confidence intervals (2,000 resamples)
- Sensitivity, specificity, precision, F1-score (at Youden’s J -optimal threshold)
- Confusion matrix

A consensus ensemble was constructed by averaging the predicted probabilities from all three models.

4 Results

4.1 Discovery Cohort: Cross-Validated Performance (Objective 1)

Table 1 summarises the 5-fold stratified CV performance on the discovery cohort.

Table 1: Discovery cohort: stratified 5-fold cross-validation AUROC.

Model	Mean AUROC	SD	Range
Random Forest	0.595	0.050	0.531-0.673
Elastic-Net	0.505	0.056	0.407-0.574
Linear SVM	0.554	0.056	0.497-0.642
Consensus Ensemble	0.532	-	-

Random Forest achieved the highest mean AUROC in cross-validation. Elastic-net’s mean near 0.5 suggests that aggressive ℓ_1 shrinkage may remove informative weak effects when n is small. Figure 1 shows discovery ROC curves.

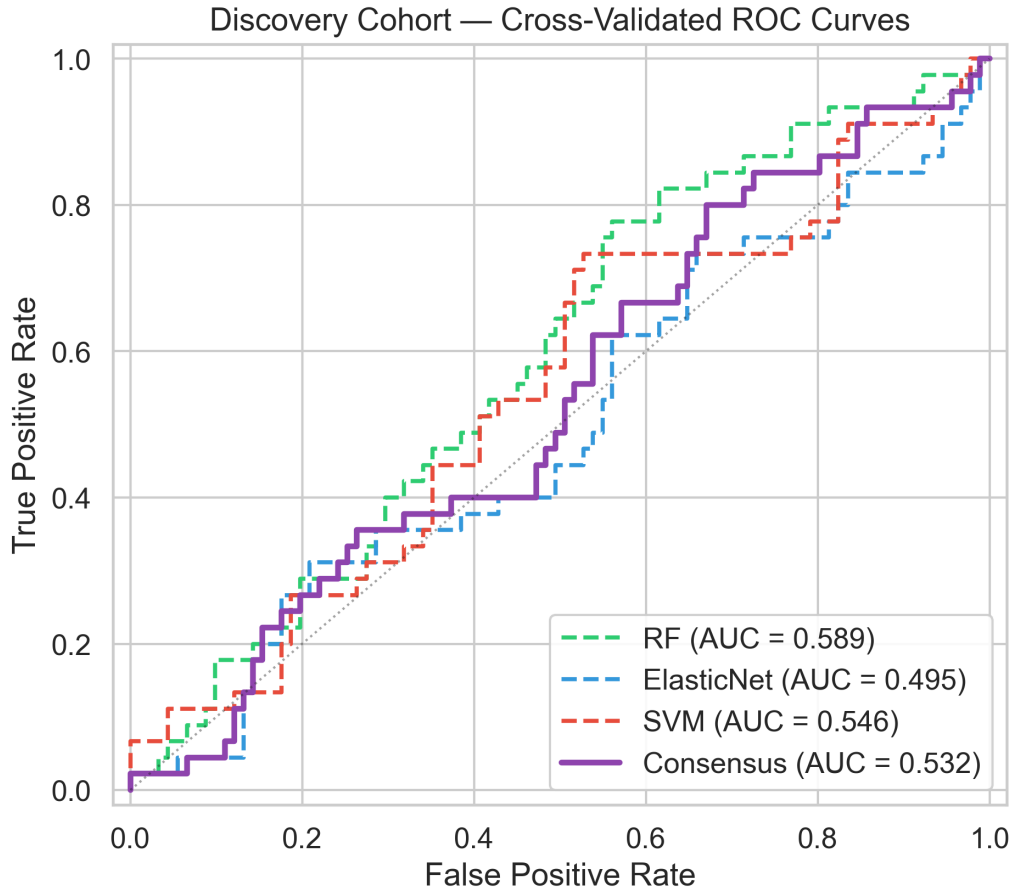
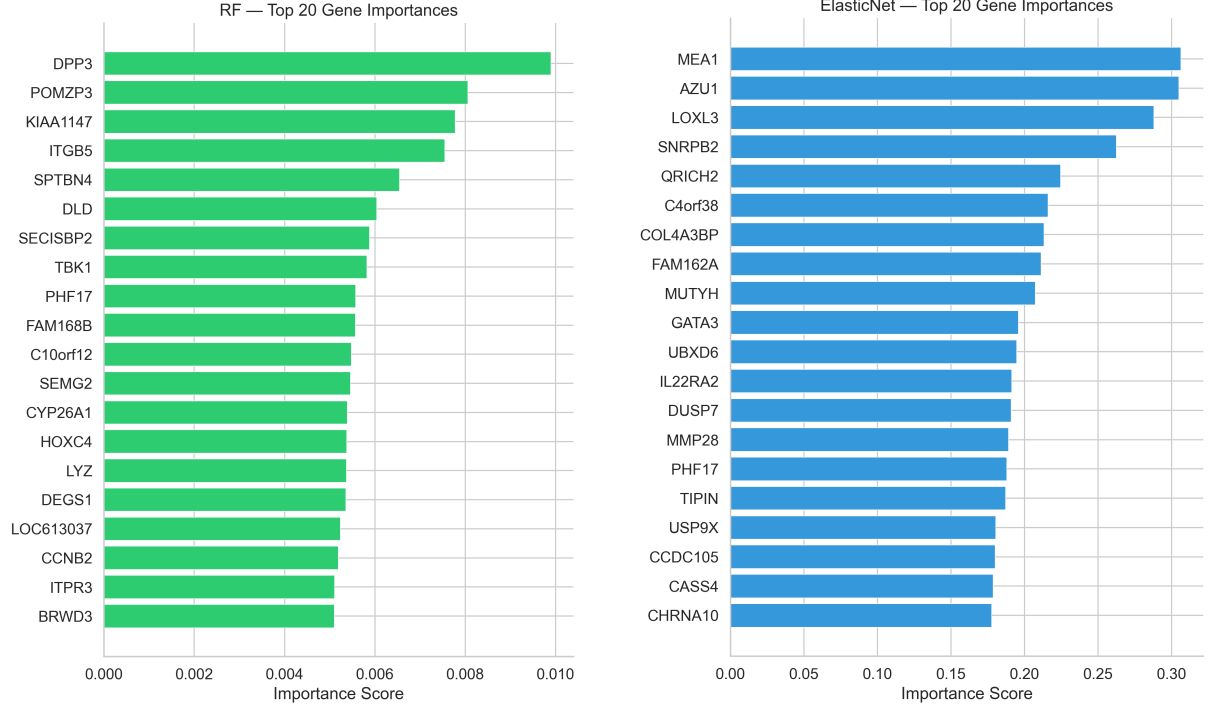


Figure 1: Discovery cohort ROC curves (cross-validated predictions).

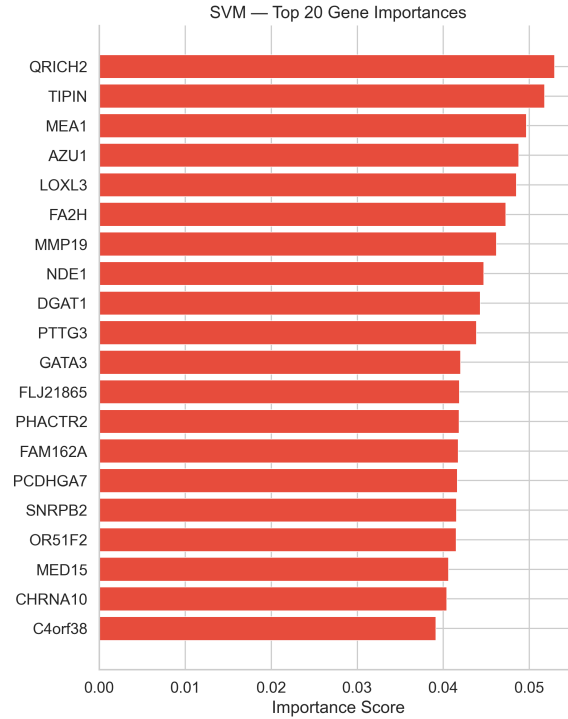
4.2 Gene-Set Stability and Feature Importance (Objective 1)

Each model identified a distinct set of top-30 genes based on its feature importance metric. Figures 2a-2c show the top-20 genes for each model.



(a) Random Forest (Gini impurity)

(b) Elastic-Net (absolute coefficients)



(c) Linear SVM (absolute weights)

Figure 2: Top-20 most important genes per model.

Top-30 rankings differed across methods, which is expected when effects are small and the feature space is large. An intersection-based consensus offers a way to down-weight idiosyncratic single-model picks (Figure 3).

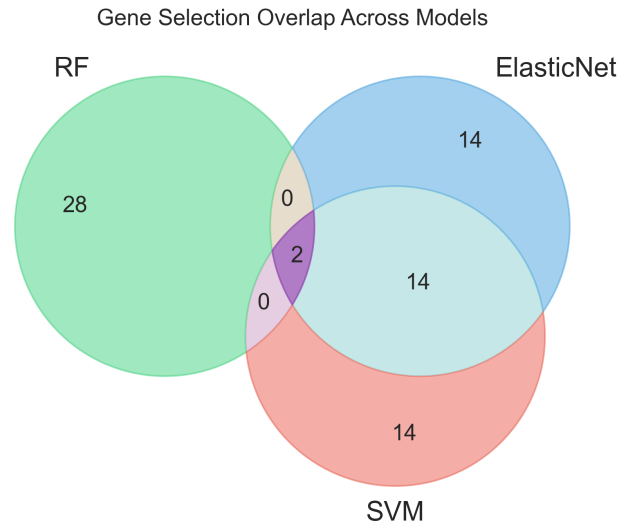


Figure 3: Venn diagram of top-30 gene overlap across the three models.

Figure 4 shows the gene stability analysis across CV folds, quantifying how consistently each gene was selected.

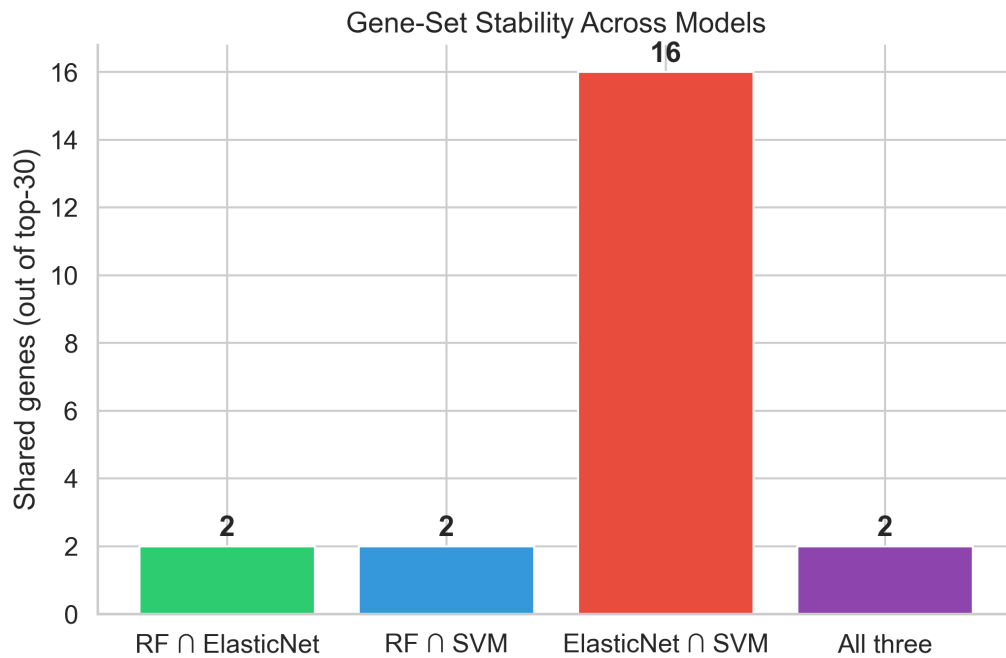


Figure 4: Gene selection stability across 5 CV folds.

4.3 Consensus Biomarker Panel (Objective 2)

The majority-vote consensus approach identified **16 genes** selected by ≥ 2 of 3 models, with **2 genes** (*AZU1*, *PHF17*) selected by all three (Table 2).

Table 2: Consensus biomarker panel: 16 genes selected by $\geq 2/3$ models.

Gene	Votes	RF	EN	SVM
<i>AZU1</i>	3	✓	✓	✓
<i>PHF17</i>	3	✓	✓	✓
<i>FLJ21865</i>	2	-	✓	✓
<i>CHRNA10</i>	2	-	✓	✓
<i>C4orf38</i>	2	-	✓	✓
<i>MEA1</i>	2	-	✓	✓
<i>LOXL3</i>	2	-	✓	✓
<i>FAM162A</i>	2	-	✓	✓
<i>CRTAM</i>	2	-	✓	✓
<i>SNRPB2</i>	2	-	✓	✓
<i>TREX1</i>	2	-	✓	✓
<i>QRICH2</i>	2	-	✓	✓
<i>PCDHGA7</i>	2	-	✓	✓
<i>USP9X</i>	2	-	✓	✓
<i>TIPIN</i>	2	-	✓	✓
<i>GATA3</i>	2	-	✓	✓

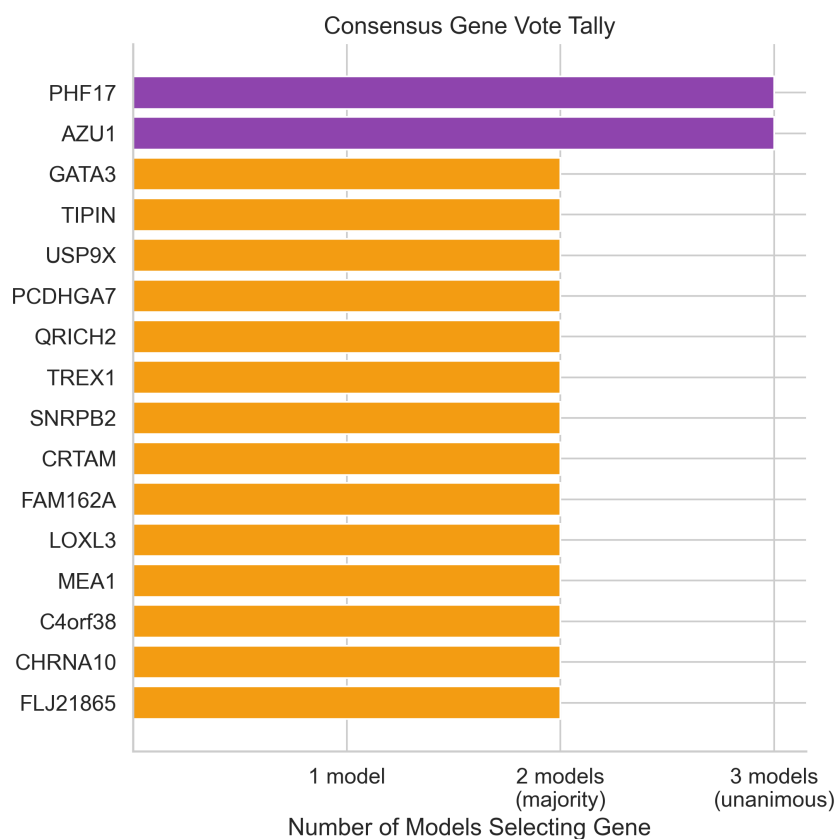


Figure 5: Gene vote tally across the three models.

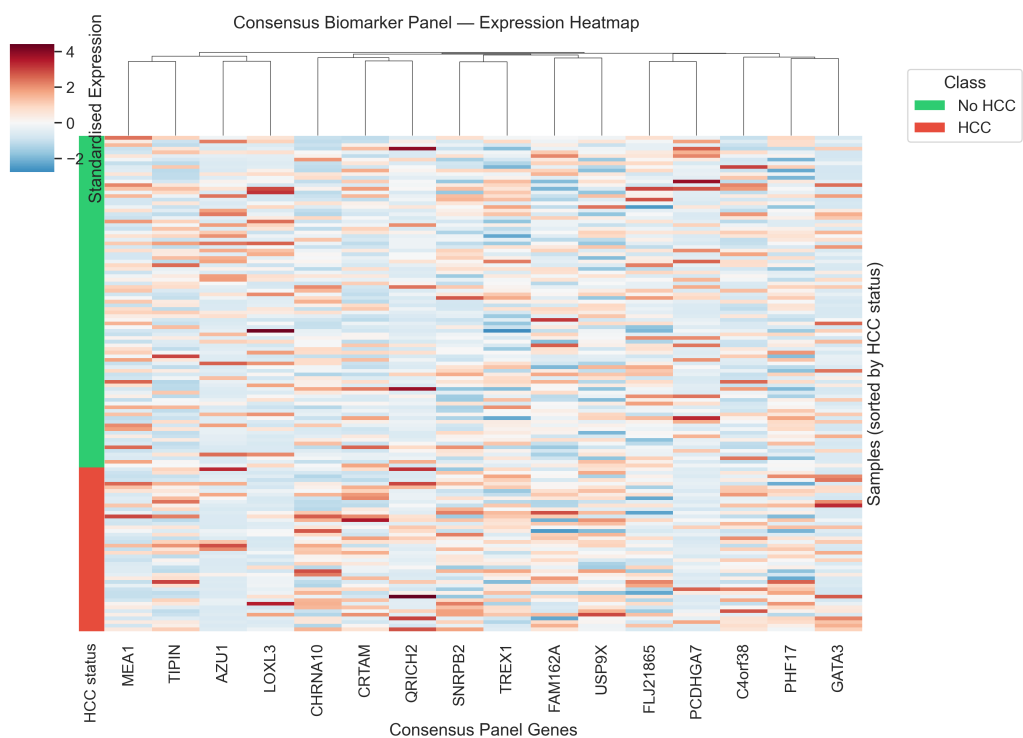


Figure 6: Expression heatmap of the 16 consensus genes across discovery samples, ordered by HCC status.

4.4 Independent Validation (Objective 3)

Table 3 reports validation cohort performance.

Table 3: Validation cohort AUROC with 95% bootstrap confidence intervals.

Model	Validation AUROC	95% CI
Random Forest	0.624	[0.456, 0.788]
Elastic-Net	0.582	[0.418, 0.738]
Linear SVM	0.606	[0.443, 0.762]
Consensus Ensemble	0.597	[0.436, 0.757]

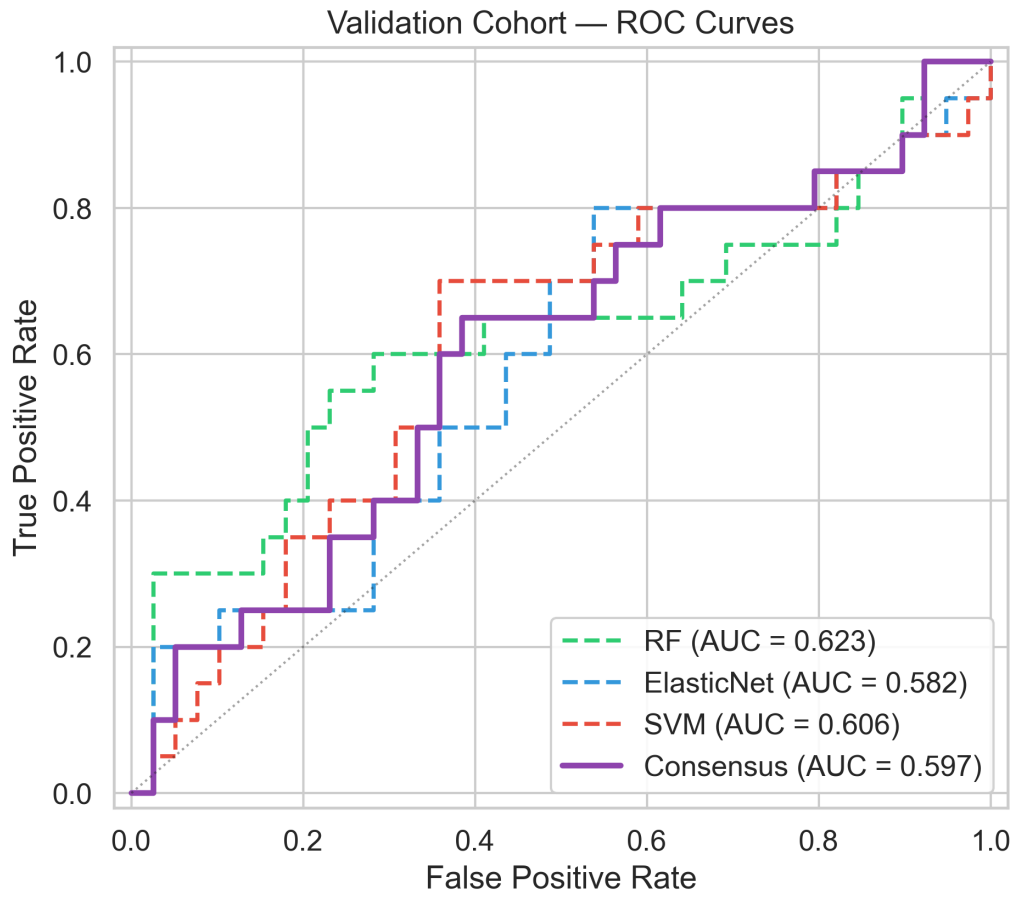


Figure 7: Validation cohort ROC curves.

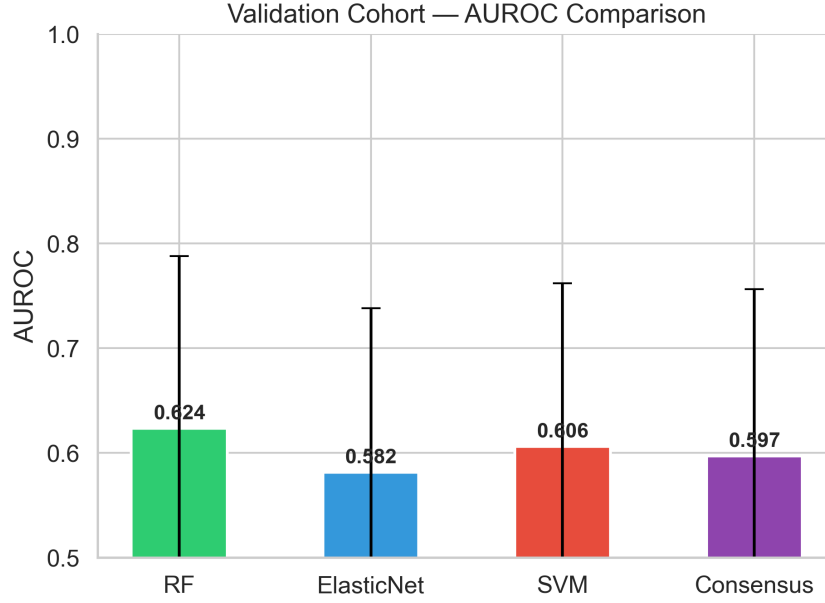


Figure 8: Validation AUROC comparison with 95% confidence intervals.

Table 4 presents the classification report for the best individual model (Random Forest) at the Youden’s J -optimal threshold.

Table 4: RF classification report on the validation cohort (threshold = 0.434).

Class	Precision	Recall	F1-Score	Support
No HCC	0.769	0.769	0.769	39
HCC	0.550	0.550	0.550	20
Accuracy	0.695			59

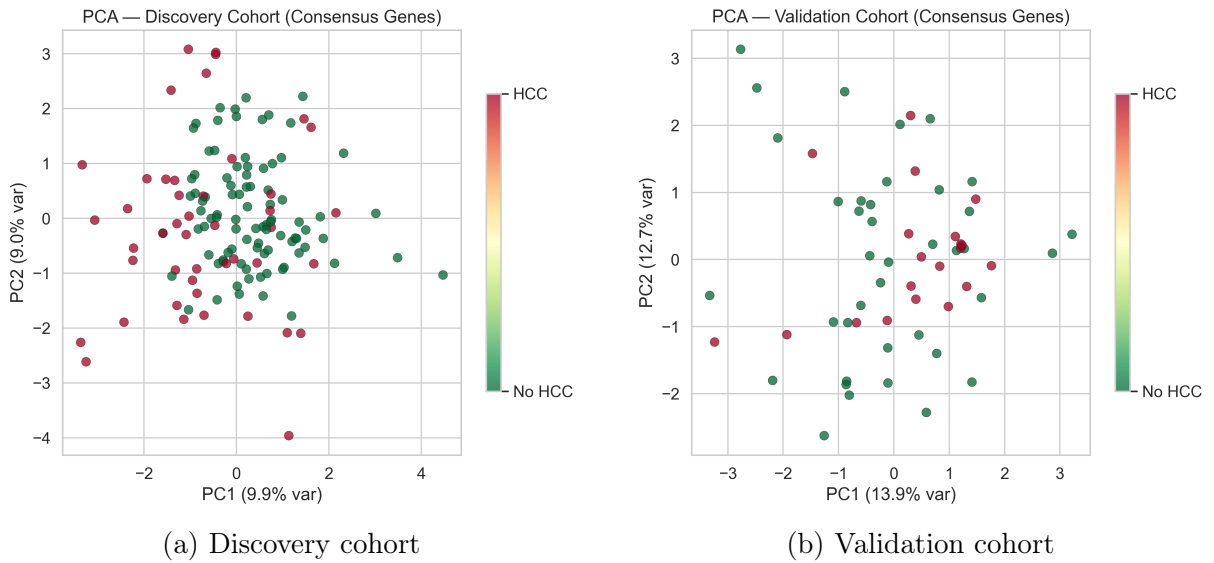


Figure 9: PCA visualisation of the 16 consensus genes, coloured by HCC outcome.

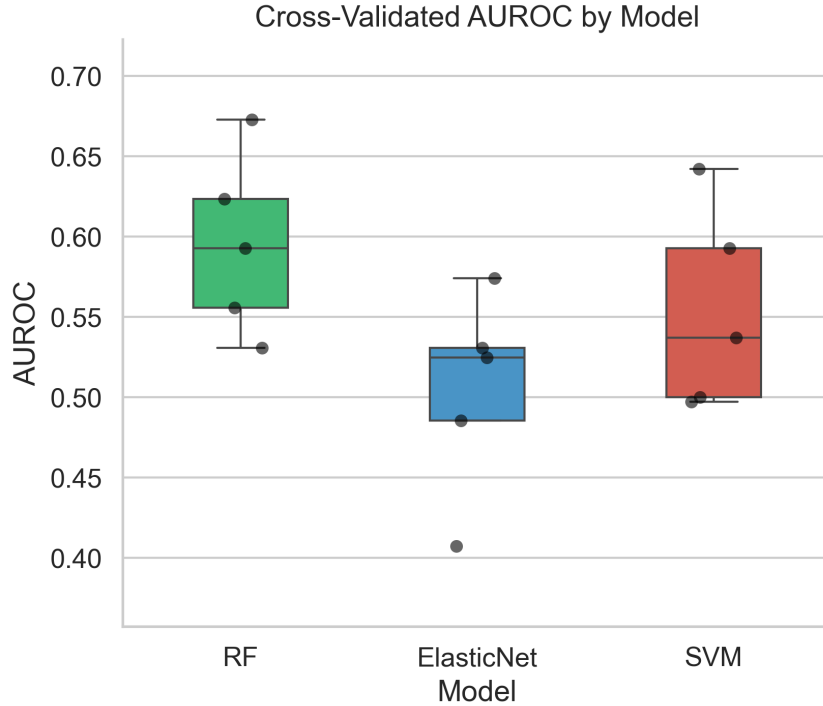


Figure 10: Cross-validation AUROC distributions across 5 folds.

5 Discussion

5.1 Performance in Context

Validation AUROCs near 0.58–0.62 sit below typical tumour-versus-normal benchmarks but are plausible for forecasting cancer from histologically benign cirrhotic tissue: the biological contrast is small, labels mix early and late events, and $p \gg n$ even after filtering. Worth noting:

1. **Sample size versus dimension.** There are 136 discovery samples and $\sim 14,700$ genes before filtering ($n/p \approx 0.009$). Even retaining 500 genes per fold leaves an ill-posed ratio.
2. **Small effects.** Expected expression shifts between future progressors and non-progressors at one biopsy time may be modest compared with overt tumour tissue.
3. **Heterogeneous time to HCC.** Events occurred from months to nearly two decades after biopsy; late cases may resemble non-cases molecularly at sampling.
4. **Single biopsy.** A cross-sectional profile cannot capture trajectory; serial sampling is rarely feasible in routine care.

Agreement between cross-validated discovery performance and the independent validation split (e.g., Random Forest ≈ 0.60 then 0.62) is consistent with the nested filtering protocol limiting overt leakage between training and evaluation.

5.2 Biological Interpretation of Consensus Genes

Two genes were selected by all three models:

- ***AZU1* (Azurocidin 1/CAP37):** A neutrophil granule protein with antimicrobial and monocyte-recruiting functions. Neutrophil infiltration and activation in cirrhotic liver has been linked to hepatocarcinogenesis through oxidative stress and NF- κ B signalling [Kim et al., 2017]. Elevated AZU1 may reflect chronic inflammatory stress that promotes malignant transformation.
- ***PHF17* (JADE1):** A component of the HBO1 histone acetyltransferase complex involved in chromatin remodelling and DNA damage response. PHF17/JADE1 has been implicated in Wnt/ β -catenin signalling regulation and has tumour-suppressive properties in renal cell carcinoma. Its dysregulation in pre-cancerous cirrhotic tissue may indicate epigenetic reprogramming preceding HCC.

Other consensus genes include *GATA3* (a transcription factor with known roles in immune cell differentiation and tumour microenvironment regulation), *LOXL3* (a lysyl oxidase family member involved in extracellular matrix remodelling during fibrosis), and *CRTAM* (a T-cell surface receptor involved in cytotoxic immune responses). Together they point to immune activation, fibrotic remodelling, and chromatin regulation as themes worth testing in follow-up work.

5.3 Comparison with Hoshida et al. (2013)

Hoshida et al. started from a published 186-gene signature learned in HCC tissue, then applied it to the cirrhotic cohort with nearest-centroid rules and survival-style analysis [Hoshida et al., 2013]. Here, genes are selected *de novo* for binary HCC outcome, and three learners are combined by vote. Limited overlap with the 186-gene list is therefore unsurprising: the learning targets and loss functions differ.

5.4 Model-Specific Behaviour

Random Forest attained the highest validation AUROC (0.624). Elastic-net lagged in both CV and validation, which may reflect that a strictly linear decision boundary with heavy penalisation underfits here. The linear SVM was close to the forest on validation (0.606), in line with its use in high dimensions when classes are only weakly separated.

Partial overlap in selected genes (Figure 3) is consistent with diffuse signal and with each method weighting features differently.

6 Limitations

1. **Small sample size.** With 195 patients (65 events), statistical power is limited. The wide confidence intervals (e.g., RF validation: 0.46-0.79) reflect this uncertainty.
2. **Internal validation only.** Both discovery and validation cohorts derive from the same centre and study. True external validation requires an independent cohort from a different institution.
3. **Pre-DAA cohort.** GSE15654 patients had active HCV cirrhosis (pre-DAA era). The transcriptomic landscape may differ in post-DAA SVR patients, though shared fibrotic and epigenetic changes likely persist.
4. **Binary classification vs. survival modelling.** We treated HCC development as binary, losing time-to-event information. Cox proportional hazards or random survival forests could better capture the temporal dimension.
5. **Single-platform limitation.** Results are specific to the Illumina HumanRef-8 v3.0 platform. Cross-platform validation (e.g., RNA-seq) would strengthen generalisability.
6. **Univariate pre-selection bias.** While nested inside CV folds, the Mann-Whitney filter still imposes a univariate assumption that may miss genes with weak individual effects but strong interactions.

7 Future Work

1. **External validation.** Validate the 16-gene consensus panel on independent HCV cirrhosis cohorts, particularly post-DAA patient cohorts with longitudinal follow-up.
2. **Survival analysis integration.** Extend the framework to incorporate time-to-HCC using Cox regression or random survival forests, leveraging the full temporal information in the data.
3. **Multi-omic integration.** Combine transcriptomic features with clinical variables (age, sex, bilirubin, platelet count, varices), methylation data, and serum biomarkers for improved risk stratification.
4. **RNA-seq validation.** Re-profile selected cohort samples using RNA sequencing to enable cross-platform validation and explore splice-variant-level features.

5. **Functional validation.** Investigate the biological roles of AZU1 and PHF17 in hepatocarcinogenesis using in vitro models or single-cell transcriptomics of cirrhotic liver tissue.

8 Conclusion

We revisited a well-characterised HCV cirrhosis microarray cohort and asked whether routine machine-learning pipelines, with pre-selection nested in cross-validation, could extract reproducible risk signal from a single liver biopsy. After training Random Forest, elastic-net, and linear SVM models and reconciling their gene rankings, we obtained held-out AUROCs in the high 0.5s to low 0.6s, and a 16-gene majority-vote set in which *AZU1* and *PHF17* were stable across all three methods.

The magnitudes are modest next to tumour-versus-control tasks, but they are in the range one might expect when the label conflates long follow-up times and the input is not yet malignant tissue. The work meets the three stated objectives: comparing learners and reporting fold stability, defining a short consensus list, and reporting validation performance with bootstrap intervals.

Code, preprocessing decisions, and figures are documented to support independent replication. The natural next step is testing the panel in external biobanks, ideally including patients observed entirely after SVR.

Acknowledgements

The author thanks Dongming Wei and Syed Hani Hassan Abidi for supervision and feedback on the project. Thanks also to Hoshida et al. for the public GSE15654 data (NCBI GEO) and to the maintainers of GEOparse. The analysis used Python 3 with Scikit-learn, NumPy, Pandas, Matplotlib, and Seaborn.

References

- Chen, X., et al. (2014). Ensemble methods for classification in bioinformatics. *Briefings in Bioinformatics*, 15(3), 398-412.
- El-Serag, H.B. (2012). Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology*, 142(6), 1264-1273.
- Falade-Nwulia, O., et al. (2017). Oral direct-acting agent therapy for hepatitis C virus infection: a systematic review. *Annals of Internal Medicine*, 166(9), 637-648.

- Hoshida, Y., et al. (2013). Prognostic gene expression signature for patients with hepatitis C-related early-stage cirrhosis. *Gastroenterology*, 144(5), 1024-1030.
- Hoshida, Y. (2014). Molecular signatures of hepatocellular carcinoma development in hepatitis C patients. *Hepatology International*, 8(S2), 443-449.
- Ioannou, G.N., et al. (2020). HCC surveillance after SVR in patients with F3/F4 fibrosis. *Journal of Hepatology*, 73(5), 1092-1099.
- Kim, J.U., et al. (2017). Immune cell phenotype associated with hepatocellular carcinoma risk. *Clinical Cancer Research*, 23(16), 4728-4738.
- Nakagawa, S., et al. (2016). Molecular liver cancer prevention in cirrhosis by organ transcriptomic analysis and lysophosphatidic acid pathway inhibition. *Cancer Cell*, 30(6), 879-890.
- Reig, M., et al. (2016). Unexpected high rate of early tumor recurrence in patients with HCV-related HCC undergoing interferon-free therapy. *Journal of Hepatology*, 65(4), 719-726.
- Sung, H., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates. *CA: A Cancer Journal for Clinicians*, 71(3), 209-249.
- Tzartzeva, K., et al. (2018). Surveillance imaging and alpha fetoprotein for early detection of hepatocellular carcinoma in patients with cirrhosis: a meta-analysis. *Gastroenterology*, 154(6), 1706-1718.

A Code Excerpts

A.1 Data Loader (Excerpt)

```

1 def load_geo_data():
2     """Load GSE15654: 195 HCV cirrhosis patients,
3     followed ~10 years. Label = HCC (1) vs no HCC (0)."""
4     expr_probes, sample_meta = _parse_series_matrix(
5         DATA_DIR / "GSE15654_series_matrix.txt.gz")
6     y_all, n_excluded = _extract_hcc_labels(
7         sample_meta, min_followup_days=5*365)
8
9     probe_to_gene = _load_gpl_annotation("GPL8432")
10    expr_genes = _probes_to_genes(expr_probes, probe_to_gene)
11
12    X_all = expr_genes.T
13    X_all = X_all.fillna(X_all.median())

```

```

14
15 X_disc, X_val, y_disc, y_val = train_test_split(
16     X_all, y_all, test_size=0.30,
17     stratify=y_all, random_state=42)
18 return X_disc, y_disc, X_val, y_val

```

Listing 1: GSE15654 data acquisition and preprocessing (data_loader.py).

A.2 Nested CV with Univariate Filter (Excerpt)

```

1 def _fold_filter_and_scale(X, y, train_idx, test_idx):
2     """Inside each CV fold: univariate filter
3     on train only, scale, apply to test."""
4     X_tr = X.iloc[train_idx]
5     X_te = X.iloc[test_idx]
6     y_tr = y.iloc[train_idx]
7
8     top_genes = univariate_filter(X_tr, y_tr,
9                                   max_genes=500)
10    X_tr = X_tr[top_genes]
11    X_te = X_te[top_genes]
12
13    scaler = StandardScaler()
14    X_tr_sc = scaler.fit_transform(X_tr)
15    X_te_sc = scaler.transform(X_te)
16    return X_tr_sc, X_te_sc, y_tr, top_genes

```

Listing 2: Univariate gene filter inside CV fold (models.py).

A.3 Consensus Derivation (Excerpt)

```

1 def derive_consensus_genes(model_results,
2                             voting_threshold=2,
3                             max_panel=20):
4     per_model_genes = {}
5     all_genes = []
6     for res in model_results:
7         per_model_genes[res["name"]] = res["top_genes"]
8         all_genes.extend(res["top_genes"])
9
10    vote_counter = Counter(all_genes)
11    consensus = [g for g, v in vote_counter.items()
12                 if v >= voting_threshold]
13    return consensus[:max_panel]

```

Listing 3: Majority-vote consensus panel (consensus.py).