

# EVW-AI

GRHECO-Xen

2025-12-17

## Contents

<b>Abstract</b>	<b>1</b>
<b>Relación entre puntuación hemorrágica y clusterización</b>	<b>8</b>
<b>Conclusión . . . . .</b>	<b>15</b>

## Abstract

El estudio aborda el problema desde dos ángulos distintos:

- Análisis fenotípico.** Se clusterizó a los pacientes basándose en variables clínicas y de laboratorio. Se identificaron 7 clusters, pero la agrupación resultó ser estadísticamente débil ( $p=0.02$ ), lo que confirma la gran dificultad de clasificar la EVW solo por sus manifestaciones fenotípicas.
- Análisis genético.** Se clusterizó a los pacientes basándose en datos genéticos (mutaciones, frecuencia, puntuaciones de patogenicidad). Se identificó un total de 9 clústeres claramente distintos y estadísticamente robustos ( $p < 0.001$ ). La conclusión más relevante es que **estos 9 clusters genéticos muestran una fuerte asociación con la severidad clínica de la enfermedad**, medida a través de la puntuación hemorrágica. Por ejemplo, el *Cluster 8* agrupa a pacientes con los tipos de VWD más severos (como 2N y 2B) y las puntuaciones hemorrágicas más altas.

### 1. Phenotypic Data Analysis

Las variables fenotípicas disponibles para el estudio son las siguientes:

```
## [1] "Centro_Familia"                  "Tipo_VWD_laboratorio_central"  
## [3] "Género"                         "Edat"  
## [5] "Peso"                            "Etnia"  
## [7] "Grupo_sanguíneo"                "Puntuación_hemorrágica"  
## [9] "FVW_Ag_LC1_resum"               "FVW_RCo_LC1_resum"  
## [11] "FVIII_C_LC1_resum"              "FVW_CB_LC1_resum"  
## [13] "FVIII_B_LC1_resum"              "FVW_FVIII_LC1_resum"  
## [15] "Análisis_multimérico_baja_LC1_resum" "Análisis_multimérico_alta_LC1_resum"  
## [17] "FVW_RCo_FVW_Ag_LC1"             "FVIII_C_FVW_Ag_LC1"
```

De las 15 disponibles, se ha decidido prescindir de 2: **Centro de Familia**, al tratarse de un identificar de centro de reclutamiento y familia, y de **FVIII\_B\_LC1**, ya que se ha comprobado que todos los valores de esta variable son 0.

**QC Phenotypic Analysis** El primer paso en nuestro análisis es realizar un QC de los datos fenotípicos. Con ello podremos comprobar si cada variable está codificada correctamente (numérica, factorial), el rango o niveles de los valores que la componen y el % de datos faltantes.

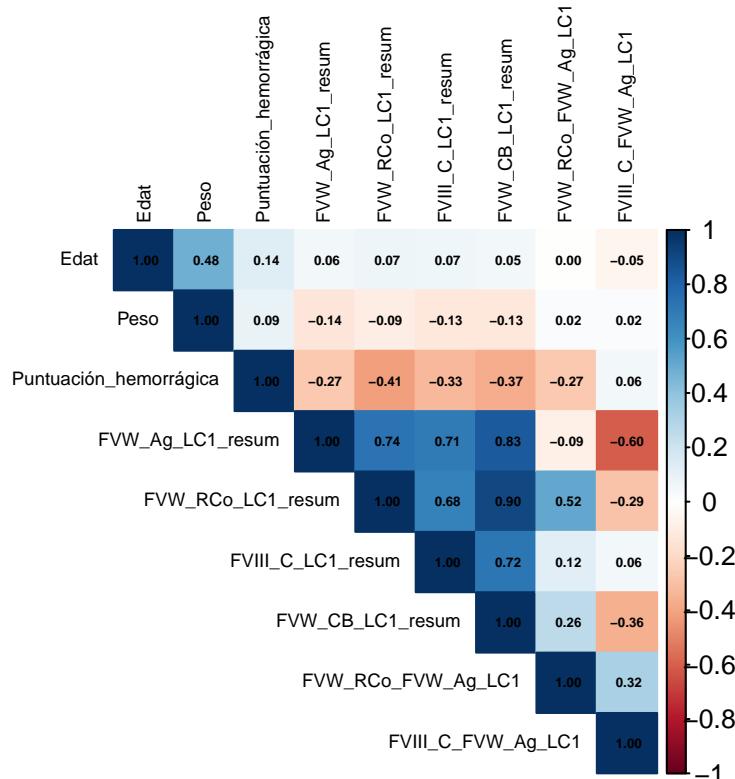
	Columna	Porcentaje_NA
## 4	Peso	45.30
## 5	Etnia	2.12
## 6	Grupo_sanguíneo	5.76
## 7	Puntuación_hemorrágica	0.61
## 8	FVW_Ag_LC1_resum	5.45
## 9	FVW_RCo_LC1_resum	11.36
## 10	FVIII_C_LC1_resum	0.91
## 11	FVW_CB_LC1_resum	6.97
## 12	FVW_FVIII_LC1_resum	94.24
## 13	Análisis_multimérico_baja_LC1_resum	4.70
## 14	Análisis_multimérico_alta_LC1_resum	27.88
## 15	FVW_RCo_FVW_Ag_LC1	12.88
## 16	FVIII_C_FVW_Ag_LC1	6.06

Se observa que la variable **FVW\_FVIII\_LC1\_resum** presenta el **94,24% de los datos faltantes**, por lo que se procede a su eliminación ya que la imputación no es la estrategia más adecuada ante tal elevado número de datos faltantes.

**Correlation analysis of Phenotypic Data** Antes de imputar los datos faltantes y explorar las relaciones entre los individuos y las variables, vamos a proceder a analizar si existe correlación entre las variables.

Se analizó la correlación entre las diferentes variables fenotípicas, mediante la correlación de Spearman, observándose que no existe ninguna variable altamente relacionada ( $>0.98$ )

## Correlaciones (spearman)



Se puede observar que no hay ninguna variable altamente correlacionada ( $\text{corr} > 0.98$ ), por lo que se emplearán las 15 variables restantes para el análisis factorial de datos mixtos (FAMD)

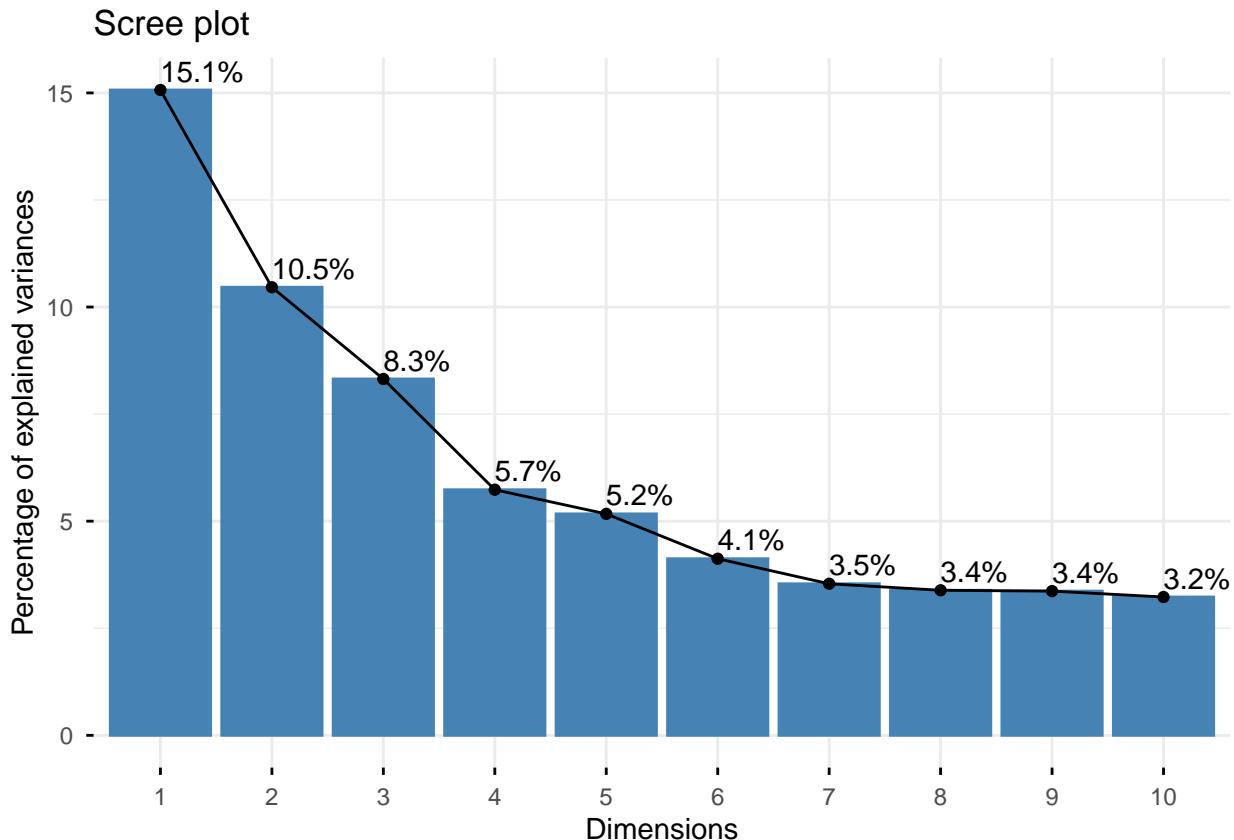
**FAMD + Clusterización** Para el análisis de los datos fenotípicos hemos decidido usar FAMD porque es una técnica estadística que permite reducir la dimensionalidad de un conjunto de datos que contiene tanto variables cuantitativas (numéricas) como cualitativas (categóricas); como es el caso de la base de datos fenotípica.

Como comentamos previamente, para hacer el FAMD, primero nos piden imputar los datos faltantes.

```
feno.imputed <- imputeFAMD(feno, ncp = 3, seed = 1923, maxiter = 1000)$completeObs
```

Una vez imputado, podemos proceder a hacer el FAMD. Para elegir el número óptimo de NCP, haremos una primera aproximación y, en base al gráfico obtenido con la primera aproximación, elegiremos el valor de NCP donde se aplana la curva. En este caso, NCP = 7

```
feno.famd <- FAMD(feno.imputed, ncp = 10, graph = FALSE)
fviz_screenplot(feno.famd, addlabels = TRUE)
```



```
feno.famd <- FAMD(feno.imputed, ncp = 7, graph = FALSE)
```

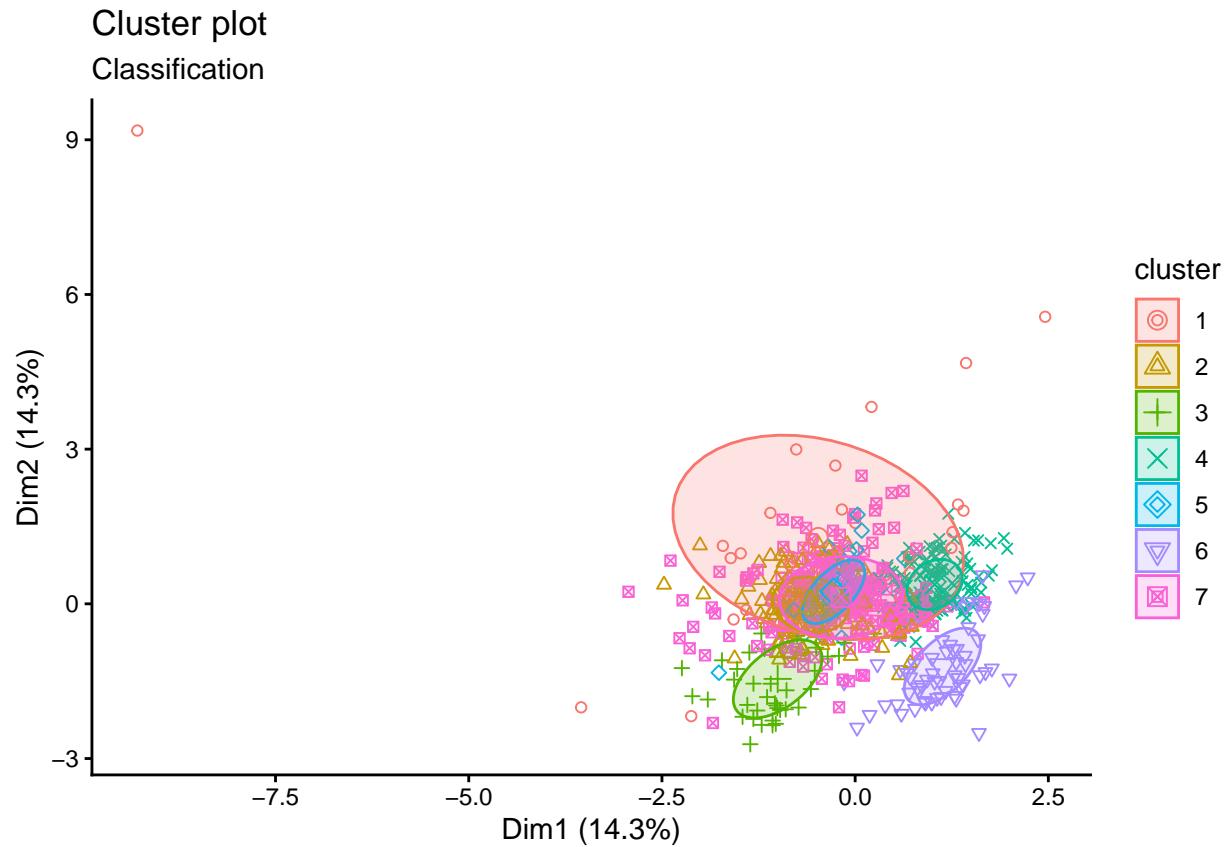
Una vez aplicado el FAMD, hemos reducido y homogeneizado los datos fenotípicos, con lo que ahora podemos clusterizar nuestras muestras.

```
# Clustering with Mclust
feno.mclust <- Mclust(feno.famd$ind$coord, G = 1:50)
```

```
##
## Número óptimo de clusters (fenotípicos): 7

##
##    1   2   3   4   5   6   7
##  39 168  40 131  40   58 184
```

Los individuos han sido clusterizados en 7 grupos diferentes, cuya distribución podemos observar en el siguiente gráfico:



Se observa que los 7 clústeres tienen un enorme solapamiento entre ellos, donde los pacientes están *mezclados* en el mismo espacio y apenas se observa una clara diferencia entre ellos. Esto muestra que los datos fenotípicos por sí solos no son un buen método para la diferenciación de pacientes. El fenotipo por sí solo es un mapa “borroso” de la enfermedad.

## 2. Genotypic Data Analysis

Hemos visto que con los datos fenotípicos no somos capaces de extraer realmente información relevante que nos permita diferenciar grupos de pacientes. Vamos a ver si con los datos genéticos podemos obtener un mapa mucho más nítido e identificar grupos de pacientes más claros.

Las variables genéticas que tenemos disponibles son las siguientes:

```
## [1] "N_var_total"          "N_var_1"           "N_var_5"
## [4] "N_alt"               "N_alt_1"           "N_alt_5"
## [7] "N_exons"              "N_dominis"         "N_aa"
## [10] "N_C2"                "N_A1"              "N_A2"
## [13] "N_A3"                "N_D3"              "N_D4"
## [16] "N_D"                 "N_D2"              "N_B"
## [19] "N_D1"                "N_C1"              "N_CK"
## [22] "N_SP"                "N_var_impact"     "N_var_frameshift"
## [25] "N_var_missense"      "N_var_Splice"      "N_var_inframe_indel"
## [28] "N_var_stop"           "N_var_ACMG"        "N_var_ClinVar"
## [31] "N_var_SIFT"            "N_var_PolyPhen"    "N_var_AlphaM"
## [34] "N_var_EVE"             "N_var_SpliceAI"    "N_var_Truncant"
## [37] "Sum_ClinVar"           "Sum_SIFT"           "Sum_PolyPhen"
```

```

## [40] "Sum_AlphaM"           "Sum_EVE"                  "Sum_Revel"
## [43] "Sum_ESAD_ESAR"        "Sum_Splice"                "Sum_Truncant"
## [46] "Sum_Impacte"          "N_aa_TOP_ClinVar"       "Score_TOP_ClinVar"
## [49] "Freq_TOP_ClinVar"      "N_aa_TOP_SIFT"           "Score_TOP_SIFT"
## [52] "Freq_TOP_SIFT"         "N_aa_TOP_PolyPhen"      "Score_TOP_PolyPhen"
## [55] "Freq_TOP_PolyPhen"     "N_aa_TOP_AlphaM"        "Score_TOP_AlphaM"
## [58] "Freq_TOP_AlphaM"       "N_aa_TOP_EVE"            "Score_TOP_EVE"
## [61] "Freq_TOP_EVE"          "N_aa_TOP_Revel"          "Score_TOP_Revel"
## [64] "Freq_TOP_Revel"        "N_aa_TOP_ESAD_ESAR"     "Score_TOP_ESAD_ESAR"
## [67] "Freq_TOP_ESAD_ESAR"    "N_aa_TOP_Splice"          "Score_TOP_Splice"
## [70] "Freq_TOP_Splice"       "N_aa_TOP_Truncant"       "Score_TOP_Truncant"
## [73] "Freq_TOP_Truncant"     "N_aa_TOP_Impacte"        "Score_TOP_Impacte"
## [76] "Freq_TOP_Impacte"      "Inv_Freq"                 "N_var_total_alt"
## [79] "N_var_5_alt"           "N_var_1_alt"              "N_exons_freq"
## [82] "N_dominis_freq"        "N_aa_freq"                "N_exons_ClinVar"
## [85] "N_dominis_ClinVar"     "N_aa_ClinVar"             "N_exons_SIFT"
## [88] "N_dominis_SIFT"         "N_aa_SIFT"                 "N_exons_PolyPhen"
## [91] "N_dominis_PolyPhen"    "N_aa_PolyPhen"            "N_exons_AlphaM"
## [94] "N_dominis_AlphaM"      "N_aa_AlphaM"               "N_exons_EVE"
## [97] "N_dominis_EVE"         "N_aa_EVE"                  "N_exons_Revel"
## [100] "N_dominis_Revel"       "N_aa_Revel"                "N_exons_Splice"
## [103] "N_dominis_Splice"      "N_aa_Splice"               "N_exons_ESAD_ESAR"
## [106] "N_dominis_ESAD_ESAR"   "N_aa_ESAD_ESAR"            "N_exons_Truncant"
## [109] "N_dominis_Truncant"    "N_aa_Truncant"             "N_exons_Impact"
## [112] "N_dominis_Impact"      "N_aa_Impact"               "N_exons_Impact"

```

**QC Genotypic Analysis** Al igual que con los datos fenotípicos, procedemos a realizar el QC de los datos genéticos para ver si está correctamente codificada cada variable, el rango de valores y el % de datos faltantes.

En el QC se observa que tenemos datos completos y no es necesaria la imputación.

```

## [1] Columna      Porcentaje_NA
## <0 rows> (or 0-length row.names)

```

En este caso, se observa que no hay datos faltantes, de manera que nos ahorraremos tener que imputar los datos. Vamos a analizar posibles correlaciones entre variables, para tratar de reducir el número de variables empleadas.

**Correlation analysis of Genotypic Data** El análisis de correlación, en este caso, sí que muestra una gran cantidad de variables con una alta correlación ( $> 0.98$ ) que se recogen en la siguiente tabla. Procedemos a la eliminación de una de las 2 variables correlacionadas.

	Variable1	Variable2	Correlacion
## 5590	N_aa_TOP_PolyPhen	N_aa_TOP_SIFT	1.0000000
## 5593	N_aa_TOP_AlphaM	N_aa_TOP_SIFT	1.0000000
## 5599	N_aa_TOP_Revel	N_aa_TOP_SIFT	1.0000000
## 5932	N_aa_TOP_AlphaM	N_aa_TOP_PolyPhen	1.0000000
## 5938	N_aa_TOP_Revel	N_aa_TOP_PolyPhen	1.0000000
## 6277	N_aa_TOP_Revel	N_aa_TOP_AlphaM	1.0000000
## 7984	N_aa_TOP_Impacte	N_aa_TOP_Truncant	0.9999903
## 12200	N_dominis_Truncant	N_exons_Truncant	0.9999341
## 644	N_var_5_alt	N_alt_5	0.9999199
## 11516	N_dominis_Splice	N_exons_Splice	0.9998025

```

## 5044 Score_TOP_Truncant      Sum_Truncant 0.9996469
## 4000      Sum_Truncant      N_var_Truncant 0.9995699
## 4027 Score_TOP_Truncant      N_var_Truncant 0.9991519
## 5081 N_dominis_Truncant      Sum_Truncant 0.9989328
## 5080 N_exons_Truncant       Sum_Truncant 0.9988878
## 305      N_var_5_alt        N_var_5      0.9988107
## 4928 Score_TOP_Splice       Sum_Splice   0.9987421
## 8132 N_dominis_Truncant      Score_TOP_Truncant 0.9986407
## 8131 N_exons_Truncant       Score_TOP_Truncant 0.9986096
## 4064 N_dominis_Truncant      N_var_Truncant 0.9984969
## 4063 N_exons_Truncant       N_var_Truncant 0.9984505
## 8133      N_aa_Truncant     Score_TOP_Truncant 0.9982405
## 4812 Score_TOP_ESAD_ESAR    Sum_ESAD_ESAR 0.9982089
## 12201 N_aa_Truncant       N_exons_Truncant 0.9982051
## 232      N_alt_5            N_var_5      0.9981925
## 12314 N_aa_Truncant       N_dominis_Truncant 0.9980705
## 5082 N_aa_Truncant       Sum_Truncant   0.9980524
## 4065 N_aa_Truncant       N_var_Truncant 0.9976965
## 532      N_var_1_alt        N_alt_1      0.9967337
## 10490 N_dominis_AlphaM     N_exons_AlphaM 0.9956768
## 9464 N_dominis_ClinVar     N_exons_ClinVar 0.9954369
## 11858 N_dominis_ESAD_ESAR N_exons_ESAD_ESAR 0.9952332
## 5596 N_aa_TOP_EVE         N_aa_TOP_SIFT  0.9949483
## 5935 N_aa_TOP_EVE         N_aa_TOP_PolyPhen 0.9949483
## 6274 N_aa_TOP_EVE         N_aa_TOP_AlphaM 0.9949483
## 6616 N_aa_TOP_Revel       N_aa_TOP_EVE  0.9949483
## 10832 N_dominis_EVE        N_exons_EVE   0.9947706
## 9806 N_dominis_SIFT        N_exons_SIFT   0.9943439
## 10148 N_dominis_PolyPhen  N_exons_PolyPhen 0.9941182
## 11174 N_dominis_Revel      N_exons_Revel  0.9937097
## 12542 N_dominis_Impact    N_exons_Impact 0.9930042
## 417      N_var_total_alt   N_alt          0.9919762
## 4963 N_aa_Splice           Sum_Splice   0.9918035
## 7788 N_aa_Splice           Score_TOP_Splice 0.9903327
## 9122 N_dominis_freq        N_exons_freq  0.9835244
## 78      N_var_total_alt   N_var_total   0.9828218

```

**Z-score + Clusterización** Al tratarse de datos cuantitativos, podemos aplicar directamente el Z-score (escalado de los datos) para normalizarlos, y, una vez normalizados, realizar el clustering. Obtenemos, del mismo modo que con los datos fenotípicos, 7 clústeres distintos:

```

geno.mclust <- Mclust(geno.scaled, G = 1:50)
feno$clusteres <- feno.mclust$classification

```

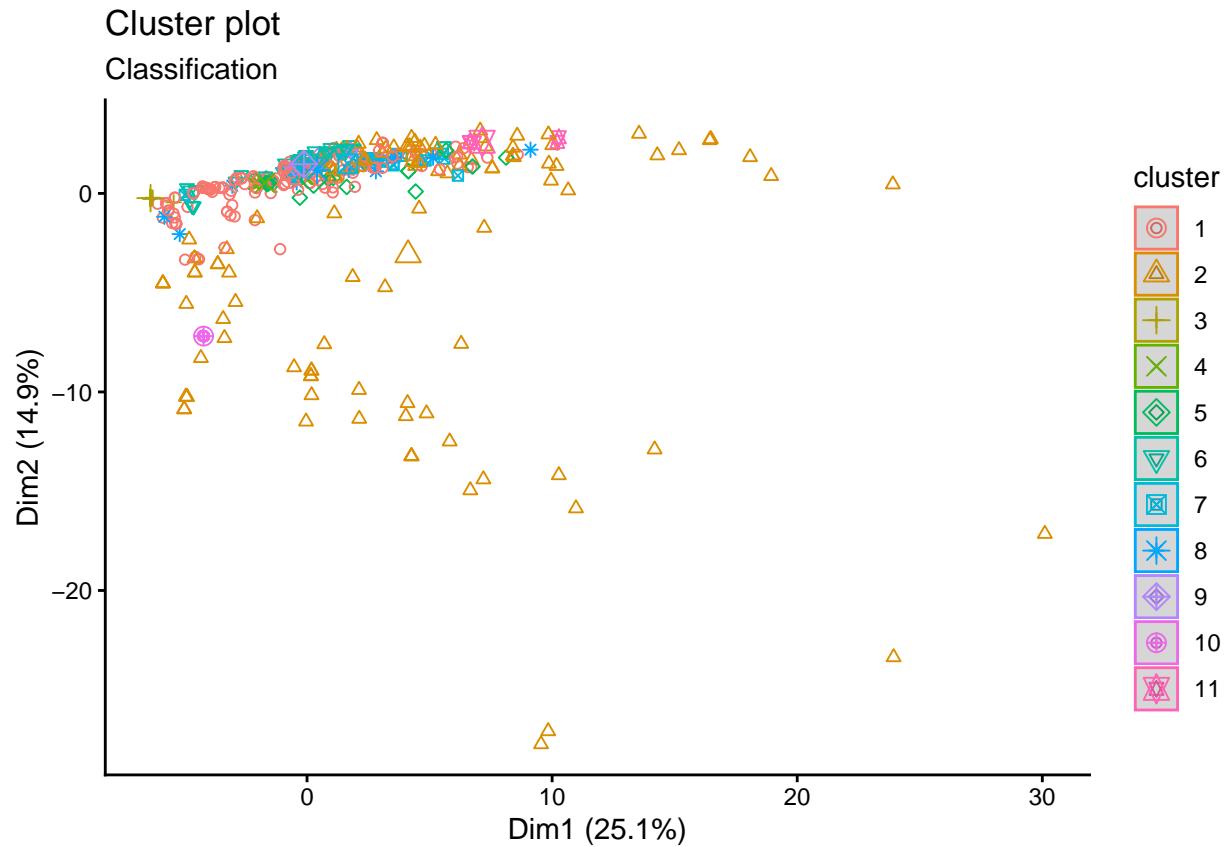
```

##
## Número óptimo de clusters (genéticos): 11

##
## 1 2 3 4 5 6 7 8 9 10 11
## 207 120 66 33 28 80 36 38 23 14 15

```

Cuya distribución en clústeres es la siguiente:



La clusterización de los datos genéticos, en cambio, sí que nos arroja alguna diferencia entre los pacientes. Hay una ligera diferenciación entre los clústeres. Se observan algunos subgrupos con más facilidad que con los datos fenotípicos.

Vamos a ver si hay alguna relación entre la puntuación hemorrágica y la clusterización.

## Relación entre puntuación hemorrágica y clusterización

Ahora que tenemos calculados el número de clústeres y a qué clúster es asignado cada individuo, procedemos al estudio de la relación entre la puntuación hemorrágica y el clúster asignado en base a las variables genéticas.

Lo primero que hacemos, es el kruskal test para ver si hay diferencias significativas entre la distribución de la puntuación hemorrágica y los clústeres. Este test responde a una pregunta general: “Si miro los 6 clusters genéticos como grupos, ¿la puntuación hemorrágica se distribuye de forma diferente entre ellos, o es la misma para todos?”

```
##  
## Kruskal-Wallis rank sum test  
##  
## data: Puntuacion_Hemorragica by geno$clusteres  
## Kruskal-Wallis chi-squared = 57.338, df = 10, p-value = 1.152e-08
```

**Se rechaza la idea de que los grupos son iguales.** Los 7 clusters tienen distribuciones de puntuación hemorrágica claramente distintas. No se trata de variación aleatoria: los clusters representan fenotipos hemorrágicos diferentes.

Con el test de Kruskal-Wallis hemos observado que “hay diferencias”, pero no dónde. El Test de Dunn es un análisis post-hoc que compara cada cluster con todos los demás (Cluster 1 vs 2, 1 vs 3) para encontrar las diferencias específicas.

```
##  Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 57.3376, df = 10, p-value = 0
##
##
##                                     Comparison of x by group
##                                     (Benjamini-Hochberg)
## Col Mean-
## Row Mean |      1      10      11      2      3      4
## -----
##      10 | -0.629208
##           | 0.3096
##           |
##      11 | 1.330070  1.424618
##           | 0.1402  0.1286
##           |
##      2 | -3.459834 -0.793726 -2.750724
##           | 0.0019*  0.2798  0.0091*
##           |
##      3 | 1.602547  1.360402 -0.451382  4.069402
##           | 0.1071  0.1365  0.3584  0.0006*
##           |
##      4 | -2.633435 -1.002827 -2.727236 -0.485880 -3.377790
##           | 0.0116*  0.2228  0.0092*  0.3519  0.0018*
##           |
##      5 | 1.554679  1.487212 -0.133137  3.385362  0.383590  3.139509
##           | 0.1100  0.1215  0.4470  0.0020*  0.3709  0.0031*
##           |
##      6 | -2.285123 -0.442941 -2.335718  0.660259 -3.170583  0.923490
##           | 0.0256  0.3547  0.0233*  0.3043  0.0030*  0.2446
##           |
##      7 | -3.397405 -1.432594 -3.175615 -1.186158 -4.051239 -0.552735
##           | 0.0021*  0.1306  0.0032*  0.1751  0.0005*  0.3325
##           |
##      8 | -2.512610 -0.862585 -2.620547 -0.243737 -3.290055  0.210871
##           | 0.0150*  0.2605  0.0115*  0.4190  0.0023*  0.4242
##           |
##      9 | -3.672652 -1.868753 -3.503886 -1.796526 -4.269338 -1.154570
##           | 0.0011*  0.0652  0.0018*  0.0738  0.0005*  0.1797
## Col Mean-
## Row Mean |      5      6      7      8
## -----
##      6 | -2.797373
##           | 0.0083*
##           |
##      7 | -3.690193 -1.591746
##           | 0.0012*  0.1057
##           |
##      8 | -3.037376 -0.715417  0.784718
```

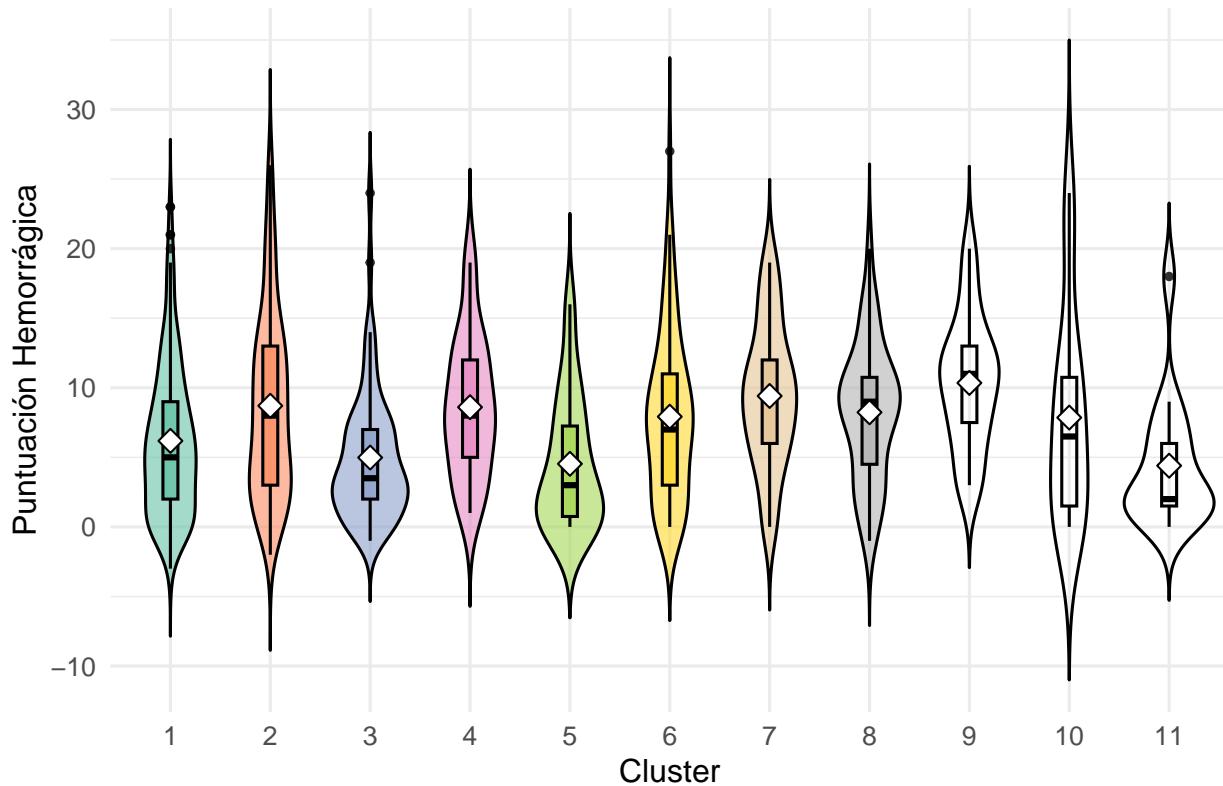
```

##          | 0.0041*   0.2965   0.2767
##          |
## 9 | -3.980903 -2.131525 -0.661314 -1.377019
##          | 0.0005*   0.0364   0.3107   0.1363
##          |
## alpha = 0.05
## Reject Ho if p <= alpha/2

```

**1. Variabilidad de la puntuación hemorrágica entre clústeres** El análisis de la puntuación hemorrágica (PH) mostró diferencias estadísticamente significativas entre los siete clústeres genéticos (Kruskal-Wallis:  $\chi^2 = 27.8$ ,  $p < 1 \times 10^{-6}$ ). Las comparaciones múltiples con corrección de Benjamini-Hochberg confirmaron que varios pares de clústeres difieren de manera consistente, indicando que los clústeres capturan diferencias fenotípicas reales.

## Distribución de la Puntuación Hemorrágica por Cluster

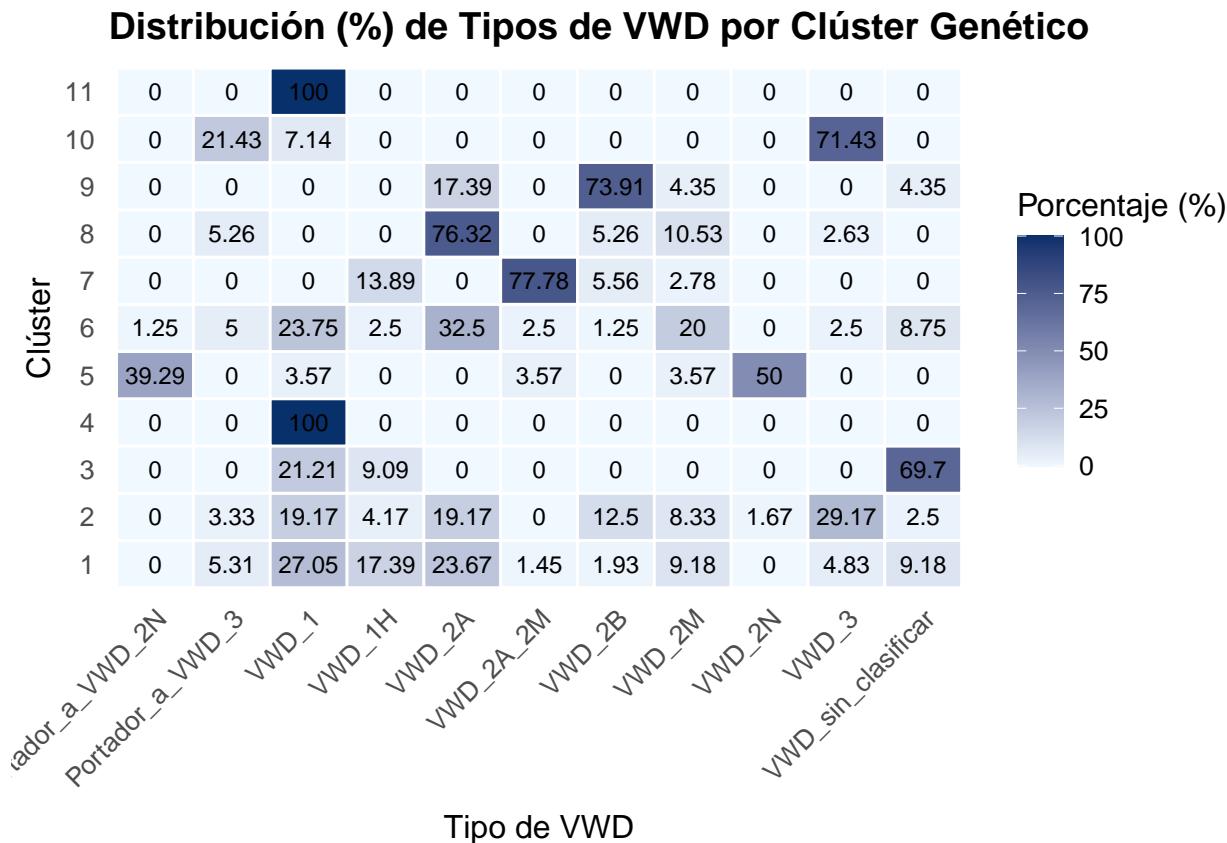


Este gráfico muestra claramente:

- Los clústeres 3 y 7 poseen las distribuciones más bajas de PH, con violines estrechos y medianas pequeñas.
- Los clústeres 2, 4 y 5 presentan niveles más altos de severidad, con medianas elevadas y densidades desplazadas hacia valores altos.
- Los clústeres 1 y 6 se sitúan en un nivel intermedio, mostrando mayor dispersión interna.

En conjunto, el patrón indica la existencia de tres niveles de severidad hemorrágica dentro de la cohorte: - clústeres de baja severidad (3 y 7), - intermedia (1, 5 y 6), - alta (2 y 4).

**2. Distribución de Tipos de VWD en cada clúster** Para comprender la composición genética de cada clúster, se analizó el porcentaje de pacientes pertenecientes a cada subtipo de VWD.



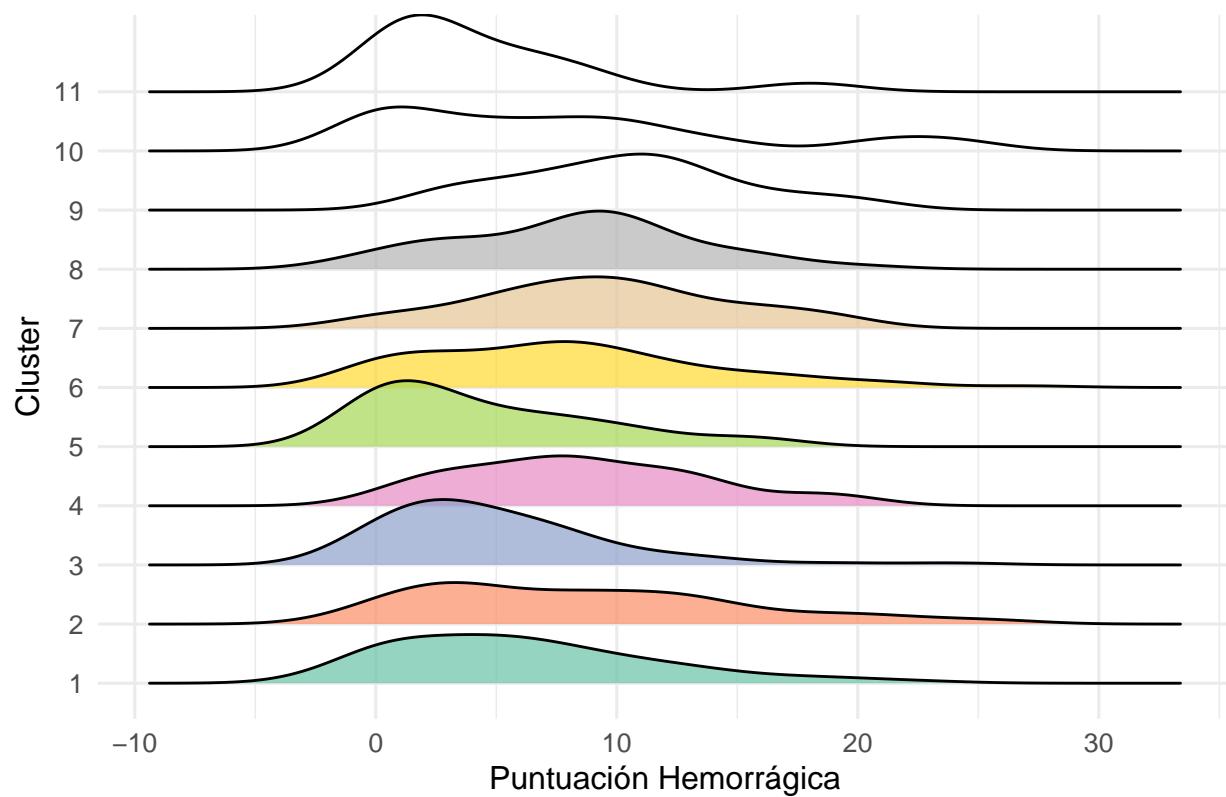
Este heatmap revela patrones muy estructurados:

- **Clúster 4 y 7 están formados íntegramente por VWD tipo 2A**, pero muestran severidades hemorrágicas opuestas, lo que ya sugiere la existencia de subgrupos moleculares contrastados dentro del mismo subtipo.
- **Clúster 5 está dominado por VWD tipo 2B (72%)**, alineándose con su PH intermedia-alta.
- **Clúster 6 se caracteriza por un predominio de VWD tipo 1H**, coherente con un fenotipo cuantitativo moderado.
- **Clústeres 1 y 2 presentan una mezcla de variantes 2A, 2B y 2N**, lo que explica su mayor variabilidad en severidad.

**3. Densidad de la Puntuación Hemorrágica** El análisis de densidad refuerza las observaciones del violin plot:

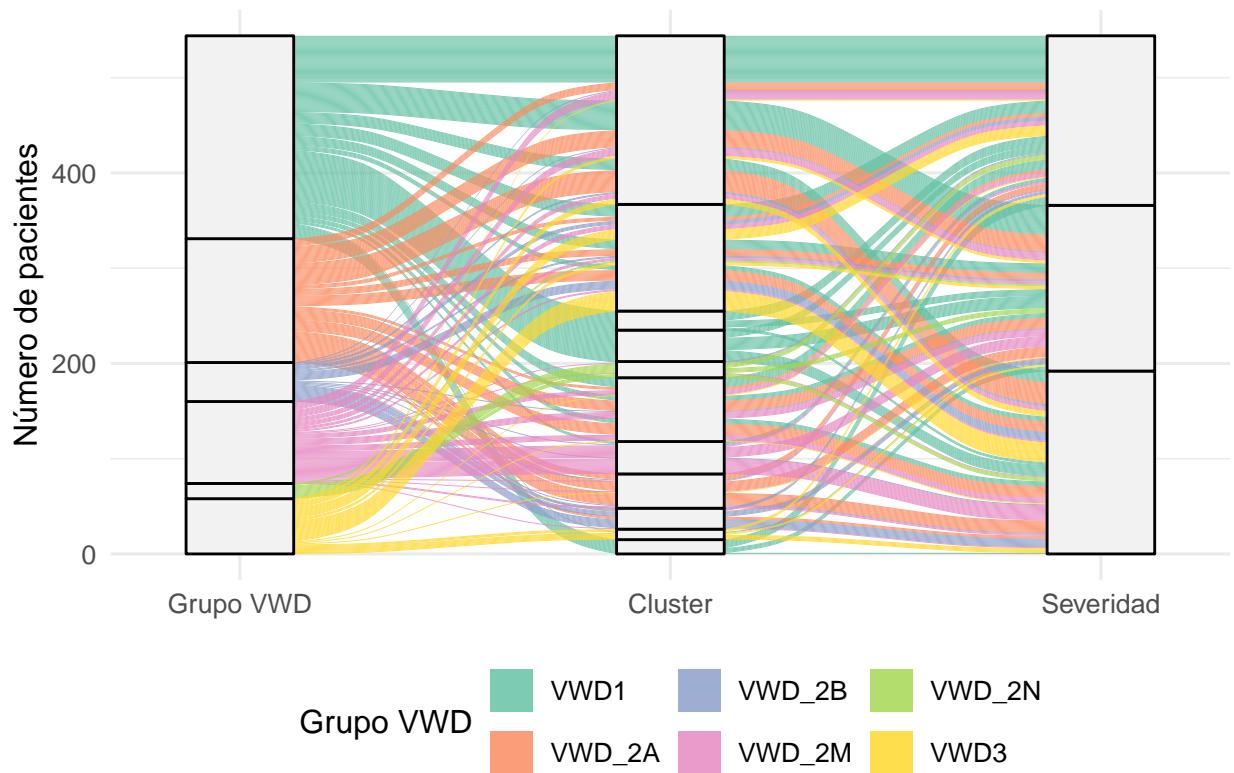
- Los **clústeres 3 y 7** muestran distribuciones muy concentradas hacia **valores bajos**.
- Los clústeres 2, 4 y 5 exhiben curvas más anchas y desplazadas hacia valores altos.
- El **clúster 1 presenta una distribución bimodal**, reflejando su mezcla de tipos de VWD funcionales.

## Distribución de Puntuación Hemorrágica por Cluster



**4. Integración Genotipo–Clúster–Severidad** Para comprender cómo se relacionan el genotipo inicial, el agrupamiento y la severidad clínica, se representaron los flujos desde el tipo de VWD hacia el clúster y posteriormente hacia la categoría de severidad.

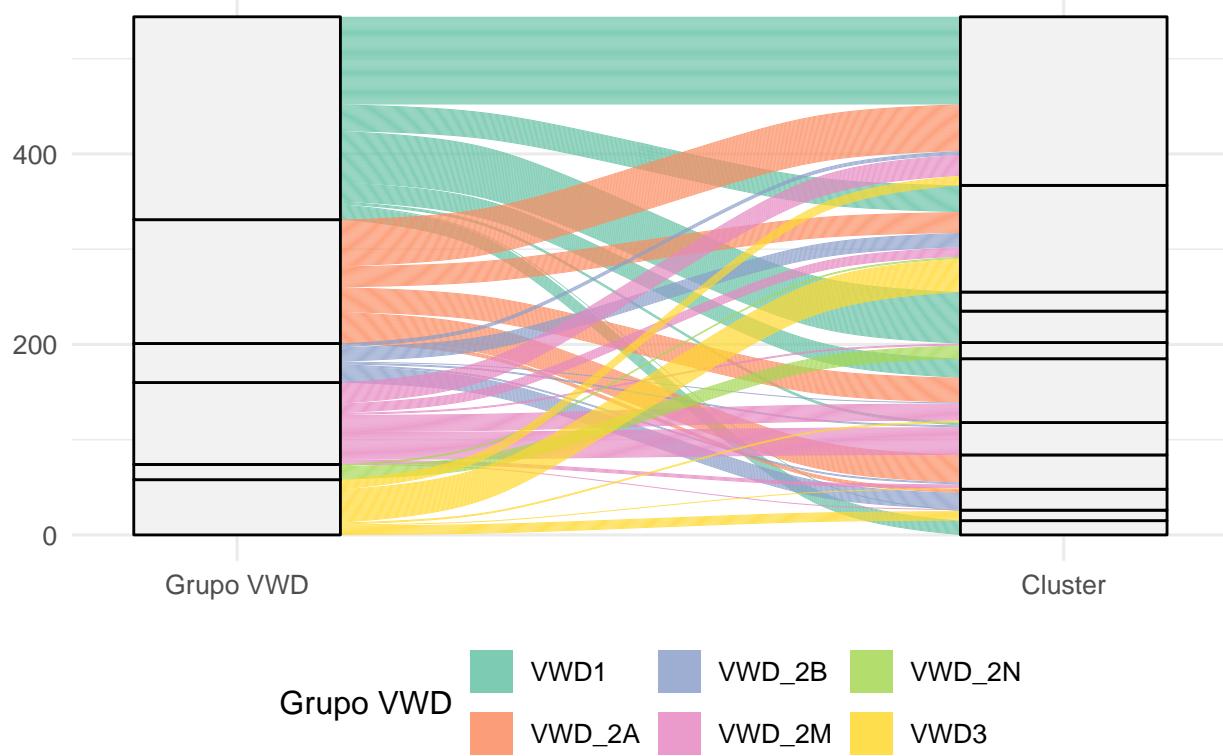
## Sankey: Grupo VWD → Cluster → Severidad Hemorrágica



Este diagrama muestra cómo los distintos tipos de VWD (especialmente los del tipo 2) se reparten entre varios clústeres, además de cómo los flujos convergen hacia distintos niveles de severidad hemorrágica y la existencia de “subtipos ocultos” dentro de categorías clásicas de VWD (por ejemplo, múltiples rutas fenotípicas dentro del tipo 2A).

También puede incluirse una versión simplificada:

## Sankey: Grupo VWD → Cluster



### 5. Interpretación integrada de cada clúster

- Clúster 1.

- **Puntuación hemorrágica:** media 7.04 (rango -3 a 27).
- **Composición VWD.** Destaca la presencia relevante de 2A (19%), 2B (21.7%) y proporciones moderadas de 2M y portadores 2N.
- **Interpretación.** Este clúster agrupa pacientes con fenotipos hemorrágicos relativamente moderados, pero genéticamente heterogéneos. La mezcla de 2A y 2B podría explicar la variabilidad interna en la severidad. No aparece una concentración clara de un subtipo específico.

- Clúster 2.

- **Puntuación hemorrágica:** media 8.65 (rango -2 a 26).
- **Composición VWD.** 2A (18.5%), 2B (18.5%), con presencia notable de 2N (12.6%) y portadores 2N (8.4%).
- **Interpretación.** Fenotípicamente algo más hemorrágico que el clúster 1. La mayor proporción de variantes funcionales (2A, 2B, 2N) se asocia bien con un fenotipo clínico algo más severo.

- Clúster 3.

- **Puntuación hemorrágica:** media 4.99 (rango -1 a 24).
- **Composición VWD.** Casi exclusivamente 2A (29%) y 2A/2M (10.5%).
- **Interpretación.** Este clúster muestra un perfil hemorrágico leve comparado con otros grupos, a pesar de incluir variantes cualitativas (2A). Podría representar un subconjunto molecular específico dentro del tipo 2A con menor impacto hemorrágico.

- Clúster 4.

- **Puntuación hemorrágica:** media 8.61 (rango 1 a 19).
- **Composición VWD.** VWD: 100% VWD tipo 2A.
- **Interpretación.** Este clúster representa un grupo homogéneo genética y fenotípicamente. La ausencia de puntuaciones muy bajas y la media alta sugieren que corresponde a un subgrupo más hemorrágico dentro del espectro 2A, probablemente asociado a variantes más severas del dominio A2.
- **Clúster 5.**
  - **Puntuación hemorrágica:** media 8.26 (rango -1 a 20).
  - **Composición VWD.** VWD: 72% VWD tipo 2B, junto con presencia de 2N y portadores 2N.
  - **Interpretación.** El clúster 5 concentra una proporción muy alta de 2B, lo que encaja con un fenotipo hemorrágico moderado-alto. La dispersión clínica puede explicarse por variabilidad funcional entre mutaciones del dominio A1.
- **Clúster 6.**
  - **Puntuación hemorrágica:** media 7.86 (rango 0 a 24).
  - **Composición VWD.** VWD: 21.4% tipo 1H, junto con 2A (7.1%).
  - **Interpretación.** Agrupa pacientes con déficit cuantitativo moderado, con puntuaciones hemorrágicas no extremas. Puede corresponder a una forma intermedia del tipo 1, probablemente influida por variantes reguladoras.
- **Clúster 7.**
  - **Puntuación hemorrágica:** media 4.40 (rango 0 a 18).
  - **Composición VWD.** VWD: 100% tipo 2A.
  - **Interpretación.** A diferencia del clúster 4, este grupo muestra baja severidad hemorrágica. Esto sugiere la existencia de subgrupos moleculares dentro del tipo 2A: uno más hemorrágico (clúster 4) y otro más benigno (clúster 7).

## Conclusión

Los clústeres identificados integran de forma coherente:

- diferencias en severidad hemorrágica,
- patrones genéticos específicos por subtipo de VWD,
- y subestructuras moleculares dentro de los propios tipos clásicos de VWD.

En conjunto, los resultados revelan que:

- **Los tipos de VWD no capturan completamente la variabilidad hemorrágica, mientras que los clústeres sí.**
- Existen **subtipos clínicos dentro del tipo 2A:** benigno (clúster 7) y severo (clúster 4).
- Las variantes tipo 2B se agrupan en un clúster bien definido (clúster 5) con severidad intermedia-alta.
- Existe un continuo clínico de severidad, representado muy bien en los violin plots y en los perfiles de densidad.

Este enfoque aporta una estratificación más fina y clínicamente relevante que la clasificación tradicional del VWD.

**Bonus Track: R2** Se ha estudiado también si existe una asociación lineal del score fenotípico + genético + el tipo de VWD respecto a la Puntuación Hemorrágica. El score fenotípico y score genético se define como el clúster asignado al paciente (en función de las variables fenotípicas y genotípicas respectivamente).

Se comprueba que no existe una relación estadística:

```
##
## -----
## #### Summary del modelo: Fenotipico ####
## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.7037 -4.5438 -0.3673  3.5144 19.4750
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  6.81481   1.11246   6.126 1.99e-09 ***
## Score_Fenotipico2 0.88889   1.21864   0.729   0.466    
## Score_Fenotipico3 -0.28981   1.43977  -0.201   0.841    
## Score_Fenotipico4  1.55253   1.25640   1.236   0.217    
## Score_Fenotipico5 -0.21481   1.48064  -0.145   0.885    
## Score_Fenotipico6  0.01852   1.53343   0.012   0.990    
## Score_Fenotipico7  0.54151   1.27344   0.425   0.671    
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.781 on 445 degrees of freedom
## Multiple R-squared:  0.01152, Adjusted R-squared:  -0.001806
## F-statistic: 0.8645 on 6 and 445 DF,  p-value: 0.5209
## 
## 
## -----
## #### Summary del modelo: Genetico ####
## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4375 -4.2374 -0.2829  3.1538 19.2105
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  6.2829    0.4490  13.995 < 2e-16 ***
## Score_Genetico2 3.1546    0.7645   4.126 4.41e-05 ***
## Score_Genetico3 -1.4934    1.0039  -1.488  0.13756  
## Score_Genetico4  2.2771    1.1946   1.906  0.05727 .  
## Score_Genetico5 -2.0607    1.3797  -1.494  0.13601
```

```

## Score_Genetico6    1.7940    0.8892    2.018  0.04425 *
## Score_Genetico7    3.5838    1.1058    3.241  0.00128 **
## Score_Genetico8    2.5633    1.1747    2.182  0.02963 *
## Score_Genetico9    3.5921    1.4548    2.469  0.01392 *
## Score_Genetico10   3.5742    2.1397    1.670  0.09554 .
## Score_Genetico11   -3.0329   2.0078   -1.511  0.13161
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.535 on 441 degrees of freedom
## Multiple R-squared:  0.1018, Adjusted R-squared:  0.08148
## F-statistic: 5.001 on 10 and 441 DF,  p-value: 6.766e-07
##
##
## -----
## #### Summary del modelo: VWD ####
## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm)
##
## Residuals:
##      Min       1Q     Median      3Q      Max
## -14.7073  -3.4074  -0.4074   2.6667  19.4500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.5500    0.8241   5.521 5.75e-08 ***
## Tipo_VWDPortador_a_VWD_2N -3.8000    2.0185  -1.883  0.06042 .  
## Tipo_VWDPortador_a_VWD_3  -2.6056    1.4792  -1.761  0.07886 .  
## Tipo_VWDVWD_1            1.7406    0.9546   1.823  0.06892 .  
## Tipo_VWDVWD_1H           -0.1426    1.2981  -0.110  0.91258  
## Tipo_VWDVWD_2A           4.7833    0.9808   4.877 1.51e-06 ***
## Tipo_VWDVWD_2B           4.9833    1.2588   3.959 8.78e-05 ***
## Tipo_VWDVWD_2A_2M        5.2682    1.2256   4.298 2.12e-05 ***
## Tipo_VWDVWD_2M           3.3288    1.2256   2.716  0.00687 ** 
## Tipo_VWDVWD_2N           1.7833    1.9228   0.927  0.35419  
## Tipo_VWDVWD_3            7.1573    1.1583   6.179 1.47e-09 ***
## ---                        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.212 on 441 degrees of freedom
## Multiple R-squared:  0.2037, Adjusted R-squared:  0.1856
## F-statistic: 11.28 on 10 and 441 DF,  p-value: < 2.2e-16
##
##
## -----
## #### Summary del modelo: Feno + Geno ####
## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm)

```

```

##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -11.5663 -4.1404 -0.3807  3.0819 18.3248
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.5808    1.1204   4.981 9.15e-07 ***
## Score_Fenotipico2     0.7718    1.1963   0.645 0.519163
## Score_Fenotipico3    -0.8570    1.4139  -0.606 0.544784
## Score_Fenotipico4     1.5942    1.2276   1.299 0.194760
## Score_Fenotipico5     0.1237    1.4594   0.085 0.932504
## Score_Fenotipico6     0.6435    1.5430   0.417 0.676825
## Score_Fenotipico7     0.5480    1.2514   0.438 0.661671
## Score_Genetico2       3.2136    0.7670   4.190 3.38e-05 ***
## Score_Genetico3      -1.4999    1.0376  -1.446 0.149029
## Score_Genetico4       2.1773    1.2057   1.806 0.071637 .
## Score_Genetico5      -2.0435    1.3984  -1.461 0.144642
## Score_Genetico6       1.8864    0.8916   2.116 0.034933 *
## Score_Genetico7       3.7250    1.1146   3.342 0.000904 ***
## Score_Genetico8       2.5996    1.1854   2.193 0.028841 *
## Score_Genetico9       3.8411    1.4914   2.576 0.010335 *
## Score_Genetico10      3.8424    2.1758   1.766 0.078099 .
## Score_Genetico11      -2.8156    2.0435  -1.378 0.168978
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.532 on 435 degrees of freedom
## Multiple R-squared:  0.1151, Adjusted R-squared:  0.08257
## F-statistic: 3.537 on 16 and 435 DF,  p-value: 5.101e-06
##
##
## -----
## #### Summary del modelo: Feno + Geno + VWD ####
## -----
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -12.1654 -3.3859 -0.4272  2.7078 18.5877
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.58101   1.46301   2.448 0.014781 *
## Score_Fenotipico2     0.60442   1.13165   0.534 0.593548
## Score_Fenotipico3    -1.03468   1.34091  -0.772 0.440765
## Score_Fenotipico4     1.51082   1.16313   1.299 0.194677
## Score_Fenotipico5    -0.04742   1.37666  -0.034 0.972536
## Score_Fenotipico6     0.45529   1.46665   0.310 0.756387
## Score_Fenotipico7     0.60053   1.18857   0.505 0.613643
## Score_Genetico2       1.47440   0.80447   1.833 0.067538 .

```

```

## Score_Genetico3      0.32050   1.25684   0.255  0.798842
## Score_Genetico4      3.07366   1.22591   2.507  0.012539 *
## Score_Genetico5      0.68148   2.46393   0.277  0.782235
## Score_Genetico6      1.74450   0.85848   2.032  0.042767 *
## Score_Genetico7      1.96769   1.91401   1.028  0.304513
## Score_Genetico8      0.64191   1.17025   0.549  0.583620
## Score_Genetico9      1.92507   1.65922   1.160  0.246608
## Score_Genetico10     1.59826   2.16911   0.737  0.461632
## Score_Genetico11     -1.87878   1.97902   -0.949 0.342985
## Tipo_VWDPortador_a_VWD_2N -4.10756   3.04363   -1.350 0.177876
## Tipo_VWDPortador_a_VWD_3  -3.05797   1.69680   -1.802 0.072224 .
## Tipo_VWDVWD_1          1.20692   1.22043   0.989  0.323262
## Tipo_VWDVWD_1H         -0.10212   1.50117   -0.068 0.945795
## Tipo_VWDVWD_2A          4.45908   1.29761   3.436  0.000648 ***
## Tipo_VWDVWD_2B          4.07211   1.66845   2.441  0.015068 *
## Tipo_VWDVWD_2A_2M       4.03997   2.13098   1.896  0.058662 .
## Tipo_VWDVWD_2M          2.74363   1.48834   1.843  0.065964 .
## Tipo_VWDVWD_2N          1.19876   3.01118   0.398  0.690754
## Tipo_VWDVWD_3           6.38556   1.52011   4.201  3.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.177 on 425 degrees of freedom
## Multiple R-squared:  0.2427, Adjusted R-squared:  0.1964
## F-statistic: 5.239 on 26 and 425 DF,  p-value: 2.831e-14
##
##
## -----
## #### Summary del modelo: Feno + VWD ####
## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm)
## 
## Residuals:
##      Min        1Q        Median        3Q        Max 
## -13.1383  -3.4378  -0.4367   2.7979  18.5213 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3.9291    1.2515   3.140  0.00181 **  
## Score_Fenotipico2 0.7976    1.1015   0.724  0.46936  
## Score_Fenotipico3 -0.9524    1.3057  -0.729  0.46612  
## Score_Fenotipico4  1.5496    1.1360   1.364  0.17324  
## Score_Fenotipico5 -0.1698    1.3438  -0.126  0.89951  
## Score_Fenotipico6  0.5236    1.4024   0.373  0.70907  
## Score_Fenotipico7  0.7065    1.1534   0.612  0.54054  
## Tipo_VWDPortador_a_VWD_2N -3.6813    2.0347  -1.809  0.07111 .  
## Tipo_VWDPortador_a_VWD_3  -2.7521    1.4827  -1.856  0.06411 .  
## Tipo_VWDVWD_1            1.7100    0.9739   1.756  0.07982 .  
## Tipo_VWDVWD_1H           -0.3183    1.3086  -0.243  0.80795  
## Tipo_VWDVWD_2A            4.7299    0.9930   4.763  2.60e-06 *** 
## Tipo_VWDVWD_2B            5.1375    1.2657   4.059  5.84e-05 *** 

```

```

## Tipo_VWDVWD_2A_2M      5.3452     1.2362    4.324 1.90e-05 ***
## Tipo_VWDVWD_2M      3.2605     1.2318    2.647  0.00841 **
## Tipo_VWDVWD_2N      1.5649     1.9292    0.811  0.41771
## Tipo_VWDVWD_3       7.1617     1.1715    6.113 2.17e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.2 on 435 degrees of freedom
## Multiple R-squared:  0.2182, Adjusted R-squared:  0.1894
## F-statistic: 7.587 on 16 and 435 DF,  p-value: 7.772e-16
##
##
## -----
## #### Summary del modelo: Geno + VWD ####
## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5481  -3.3290  -0.4841   2.7776  19.5159
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.0694    1.1815   3.444 0.000629 ***
## Score_Genetico2 1.4191    0.8047   1.763 0.078533 .
## Score_Genetico3 0.4147    1.2257   0.338 0.735273
## Score_Genetico4 3.2310    1.2169   2.655 0.008224 ** 
## Score_Genetico5 0.4801    2.4596   0.195 0.845325
## Score_Genetico6 1.6520    0.8578   1.926 0.054779 .
## Score_Genetico7 1.8503    1.9036   0.972 0.331592
## Score_Genetico8 0.6300    1.1632   0.542 0.588381
## Score_Genetico9 1.6774    1.6316   1.028 0.304483
## Score_Genetico10 1.3860    2.1227   0.653 0.514147
## Score_Genetico11 -2.0790   1.9412  -1.071 0.284780
## Tipo_VWDPortador_a_VWD_2N -3.9460   3.0306  -1.302 0.193591
## Tipo_VWDPortador_a_VWD_3 -2.8406   1.6871  -1.684 0.092965 .
## Tipo_VWDVWD_1        1.2596   1.2103   1.041 0.298584
## Tipo_VWDVWD_1H       0.1337   1.4946   0.089 0.928787
## Tipo_VWDVWD_2A       4.5898   1.2926   3.551 0.000426 ***
## Tipo_VWDVWD_2B       4.1139   1.6658   2.470 0.013910 *
## Tipo_VWDVWD_2A_2M    4.1202   2.1273   1.937 0.053420 .
## Tipo_VWDVWD_2M       2.9151   1.4773   1.973 0.049107 *
## Tipo_VWDVWD_2N       1.6795   2.9928   0.561 0.574966
## Tipo_VWDVWD_3        6.4787   1.5096   4.292 2.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.188 on 431 degrees of freedom
## Multiple R-squared:  0.2289, Adjusted R-squared:  0.1931
## F-statistic: 6.398 on 20 and 431 DF,  p-value: 3.178e-15

```

```

##          R2   R2_Ajustado
## Fenotipico    0.01152227 -0.001805524
## Genetico      0.10184660  0.081480309
## VWD           0.20369359  0.185636752
## Feno + Geno   0.11511429  0.082566773
## Feno + Geno + VWD 0.24269824  0.196369196
## Feno + VWD    0.21816955  0.189412566
## Geno + VWD    0.22891881  0.193137778

```

Tampoco si eliminamos a los portadores o a los individuos con un VWD sin clasificar:

```

##
## -----
## #### Summary del modelo: Fenotipico ####
## -----
##
##
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm_filtrado)
##
## Residuals:
##       Min        1Q     Median        3Q       Max
## -10.0080  -4.5066  -0.4691   3.5275  19.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.6087    1.2063   6.307 7.43e-10 ***
## Score_Fenotipico2 0.3993    1.3126   0.304   0.761    
## Score_Fenotipico3 -0.9771    1.5284  -0.639   0.523    
## Score_Fenotipico4  0.8638    1.3502   0.640   0.523    
## Score_Fenotipico5 -0.7024    1.5815  -0.444   0.657    
## Score_Fenotipico6  0.3004    1.7252   0.174   0.862    
## Score_Fenotipico7 -0.1396    1.3669  -0.102   0.919    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.785 on 405 degrees of freedom
## Multiple R-squared:  0.00975,  Adjusted R-squared:  -0.00492 
## F-statistic: 0.6646 on 6 and 405 DF,  p-value: 0.6783
##
##
## -----
## #### Summary del modelo: Genetico ####
## -----
##
##
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm_filtrado)
##
## Residuals:
##       Min        1Q     Median        3Q       Max
## -9.867  -4.260  -0.375   3.418  16.418
##
## Coefficients:

```

```

##                                     Estimate Std. Error t value Pr(>|t|) 
## (Intercept)                 6.3750    0.4641 13.735 < 2e-16 ***
## Score_Genetico2            3.2073    0.7798  4.113 4.74e-05 ***
## Score_Genetico3           -0.3750    1.7423 -0.215  0.82969
## Score_Genetico4            2.1850    1.2068  1.811  0.07095 .
## Score_Genetico5           -2.1528    1.3924 -1.546  0.12288
## Score_Genetico6            1.7292    0.9283  1.863  0.06323 .
## Score_Genetico7            3.4917    1.1178  3.124  0.00192 **
## Score_Genetico8            2.4712    1.1868  2.082  0.03796 *
## Score_Genetico9            3.5000    1.4677  2.385  0.01756 *
## Score_Genetico10           3.4821    2.1557  1.615  0.10703
## Score_Genetico11           3.1250    2.0231 -1.545  0.12323
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.57 on 401 degrees of freedom
## Multiple R-squared:  0.09121, Adjusted R-squared:  0.06855 
## F-statistic: 4.025 on 10 and 401 DF, p-value: 2.83e-05
## 
## -----
## #### Summary del modelo: VWD ####
## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm_filtrado)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -14.7073  -3.3519 -0.3333  2.7094 16.5926 
## 
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|) 
## (Intercept)                  0.750     1.849   0.406  0.685315
## Tipo_VWDPortador_a_VWD_3   1.194     2.223   0.537  0.591320
## Tipo_VWDVWD_1               5.541     1.912   2.898  0.003958 ** 
## Tipo_VWDVWD_1H              3.657     2.106   1.737  0.083176 . 
## Tipo_VWDVWD_2A              8.583     1.925   4.459  1.07e-05 *** 
## Tipo_VWDVWD_2B              8.783     2.082   4.220  3.03e-05 *** 
## Tipo_VWDVWD_2A_2M           9.068     2.062   4.399  1.40e-05 *** 
## Tipo_VWDVWD_2M              7.129     2.062   3.458  0.000602 *** 
## Tipo_VWDVWD_2N              5.583     2.542   2.197  0.028625 *  
## Tipo_VWDVWD_3                10.957    2.022   5.419  1.03e-07 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.231 on 402 degrees of freedom
## Multiple R-squared:  0.1963, Adjusted R-squared:  0.1783 
## F-statistic: 10.91 on 9 and 402 DF, p-value: 3.09e-15
## 
## -----
## #### Summary del modelo: Feno + Geno ####

```

```

## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm_filtrado)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -9.7215 -4.2376 -0.3986  3.3319 18.0108 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  6.1988    1.2239   5.065 6.29e-07 ***
## Score_Fenotipico2  0.3074    1.3020   0.236  0.81348  
## Score_Fenotipico3 -1.5107    1.5182  -0.995  0.32030  
## Score_Fenotipico4  0.9143    1.3316   0.687  0.49273  
## Score_Fenotipico5 -0.4372    1.5705  -0.278  0.78088  
## Score_Fenotipico6  0.2861    1.7345   0.165  0.86906  
## Score_Fenotipico7 -0.2092    1.3550  -0.154  0.87738  
## Score_Genetico2   3.3012    0.7834   4.214 3.11e-05 ***
## Score_Genetico3   -0.2407   1.7500  -0.138  0.89067  
## Score_Genetico4   2.1021    1.2202   1.723  0.08571 .  
## Score_Genetico5   -2.0753   1.4145  -1.467  0.14312  
## Score_Genetico6   1.8924    0.9347   2.025  0.04359 *  
## Score_Genetico7   3.7138    1.1286   3.291  0.00109 ** 
## Score_Genetico8   2.6084    1.1982   2.177  0.03007 *  
## Score_Genetico9   3.7732    1.5134   2.493  0.01307 *  
## Score_Genetico10  3.5762    2.2020   1.624  0.10515  
## Score_Genetico11 -2.9097   2.0645  -1.409  0.15950  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.571 on 395 degrees of freedom
## Multiple R-squared:  0.1043, Adjusted R-squared:  0.06799 
## F-statistic: 2.874 on 16 and 395 DF,  p-value: 0.0001801
## 
## 
## ----- 
## #### Summary del modelo: Feno + Geno + VWD ####
## ----- 
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm_filtrado)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.8895 -3.1951 -0.3743  2.7977 15.9642 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.1478    3.0636  -0.048  0.96154  
## Score_Fenotipico2  0.3057    1.2262   0.249  0.80328  
## Score_Fenotipico3 -1.4992    1.4340  -1.045  0.29645 

```

```

## Score_Fenotipico4      1.0632    1.2540    0.848   0.39706
## Score_Fenotipico5     -0.3625    1.4725   -0.246   0.80568
## Score_Fenotipico6      0.2528    1.6486    0.153   0.87822
## Score_Fenotipico7      0.1254    1.2763    0.098   0.92176
## Score_Genetico2       1.6565    0.8183    2.024   0.04362 *
## Score_Genetico3       1.0042    1.6781    0.598   0.54989
## Score_Genetico4       3.1395    1.2392    2.533   0.01169 *
## Score_Genetico5       0.7229    2.4793    0.292   0.77076
## Score_Genetico6       1.6634    0.8997    1.849   0.06525 .
## Score_Genetico7       2.0374    1.9274    1.057   0.29114
## Score_Genetico8       0.7241    1.1788    0.614   0.53939
## Score_Genetico9       2.0210    1.6769    1.205   0.22888
## Score_Genetico10      1.5540    2.1898    0.710   0.47835
## Score_Genetico11      -1.8446    1.9956   -0.924   0.35588
## Tipo_VWDPortador_a_VWD_3 1.0192    3.0948    0.329   0.74209
## Tipo_VWDVWD_1          5.2165    2.8739    1.815   0.07028 .
## Tipo_VWDVWD_1H         3.9102    3.0170    1.296   0.19572
## Tipo_VWDVWD_2A         8.5115    2.8882    2.947   0.00340 **
## Tipo_VWDVWD_2B         8.0263    3.0512    2.631   0.00887 **
## Tipo_VWDVWD_2A_2M      8.0949    3.1548    2.566   0.01067 *
## Tipo_VWDVWD_2M          6.8279    2.9106    2.346   0.01949 *
## Tipo_VWDVWD_2N          5.2082    2.5671    2.029   0.04316 *
## Tipo_VWDVWD_3           10.3579   2.9851    3.470   0.00058 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.201 on 386 degrees of freedom
## Multiple R-squared:  0.2372, Adjusted R-squared:  0.1878
## F-statistic: 4.801 on 25 and 386 DF,  p-value: 3.267e-12
##
##
## -----
## #### Summary del modelo: Feno + VWD ####
## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm_filtrado)
## 
## Residuals:
##      Min       1Q       Median      3Q      Max 
## -13.0554  -3.4860  -0.4975   2.8089  15.9923 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.5850    2.1727   0.269   0.787887    
## Score_Fenotipico2 0.5702    1.1897   0.479   0.631987    
## Score_Fenotipico3 -1.3270    1.3894  -0.955   0.340117    
## Score_Fenotipico4  1.1395    1.2244   0.931   0.352592    
## Score_Fenotipico5 -0.4231    1.4413  -0.294   0.769276    
## Score_Fenotipico6  0.4468    1.5877   0.281   0.778558    
## Score_Fenotipico7  0.3221    1.2416   0.259   0.795430    
## Tipo_VWDPortador_a_VWD_3 0.8825    2.2390   0.394   0.693692    
## Tipo_VWDVWD_1        5.3505    1.9230   2.782   0.005655 ** 

```

```

## Tipo_VWDVWD_1H          3.3175    2.1143    1.569 0.117428
## Tipo_VWDVWD_2A          8.3661    1.9356    4.322 1.96e-05 ***
## Tipo_VWDVWD_2B          8.7497    2.0890    4.188 3.46e-05 ***
## Tipo_VWDVWD_2A_2M       8.9974    2.0670    4.353 1.71e-05 ***
## Tipo_VWDVWD_2M          6.9043    2.0690    3.337 0.000927 ***
## Tipo_VWDVWD_2N          5.1675    2.5747    2.007 0.045429 *
## Tipo_VWDVWD_3           10.7974   2.0350    5.306 1.87e-07 ***
##
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.225 on 396 degrees of freedom
## Multiple R-squared: 0.2101, Adjusted R-squared: 0.1802
## F-statistic: 7.023 on 15 and 396 DF, p-value: 1.195e-13
##
## -----
## #### Summary del modelo: Geno + VWD ####
## -----
## 
## 
## Call:
## lm(formula = modelos[[nombre]], data = dataframe_glm_filtrado)
##
## Residuals:
##      Min       1Q     Median      3Q      Max
## -13.4404  -3.2416  -0.2958  2.8830  15.2736
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.1402    2.8552   0.049 0.960874    
## Score_Genetico2 1.5724    0.8176   1.923 0.055166 .  
## Score_Genetico3 0.9338    1.6720   0.558 0.576832    
## Score_Genetico4 3.2911    1.2285   2.679 0.007697 ** 
## Score_Genetico5 0.4831    2.4709   0.196 0.845101    
## Score_Genetico6 1.4972    0.8953   1.672 0.095245 .  
## Score_Genetico7 1.8312    1.9123   0.958 0.338872    
## Score_Genetico8 0.6380    1.1697   0.545 0.585757    
## Score_Genetico9 1.7304    1.6397   1.055 0.291917    
## Score_Genetico10 1.4710    2.1322   0.690 0.490673    
## Score_Genetico11 -2.0189   1.9528  -1.034 0.301824    
## Tipo_VWDPortador_a_VWD_3 1.0923    3.0758   0.355 0.722674    
## Tipo_VWDVWD_1            5.1288    2.8613   1.792 0.073831 .  
## Tipo_VWDVWD_1H           4.0138    3.0108   1.333 0.183266    
## Tipo_VWDVWD_2A           8.5176    2.8760   2.962 0.003247 ** 
## Tipo_VWDVWD_2B           7.9710    3.0437   2.619 0.009166 ** 
## Tipo_VWDVWD_2A_2M        8.0744    3.1518   2.562 0.010785 *  
## Tipo_VWDVWD_2M           6.8661    2.8993   2.368 0.018360 *  
## Tipo_VWDVWD_2N           5.5891    2.5330   2.206 0.027930 *  
## Tipo_VWDVWD_3            10.3002   2.9711   3.467 0.000585 *** 
##
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.206 on 392 degrees of freedom
## Multiple R-squared: 0.2237, Adjusted R-squared: 0.1861

```

```

## F-statistic: 5.946 on 19 and 392 DF, p-value: 3.031e-13

##          R2   R2_Ajustado
## Fenotipico 0.009749969 -0.004920402
## Genetico    0.091209373  0.068546265
## VWD        0.196327880  0.178335221
## Feno + Geno 0.104276417  0.067993942
## Feno + Geno + VWD 0.237209598  0.187806075
## Feno + VWD   0.210126109  0.180206644
## Geno + VWD   0.223709923  0.186083618

```

Para evaluar la contribución relativa de los diferentes predictores categóricos (score fenotípico, score genético y tipo de VWD) sobre la variabilidad de la Puntuación Hemorrágica, se realizaron comparaciones jerárquicas entre modelos lineales mediante análisis de varianza (ANOVA).

Este enfoque permite determinar si la incorporación de un nuevo bloque de variables mejora significativamente el ajuste del modelo en comparación con un modelo más simple que contiene únicamente un subconjunto de predictores. En otras palabras, ANOVA evalúa si las nuevas variables explican una proporción adicional de la variabilidad del fenotipo que no estaba capturada por los predictores previamente incluidos.

Para ello, se definieron múltiples modelos anidados que representan distintas combinaciones de los predictores de interés. Cada par de modelos fue comparado mediante ANOVA tipo I, obteniéndose valores F y p que indican si las diferencias en la devianza residual entre modelos son estadísticamente significativas. De esta forma, se puede determinar:

- si los clústeres fenotípicos o genéticos aportan información no redundante;
- si el tipo de VWD añade capacidad explicativa más allá de los clústeres;
- y qué predictores contribuyen de forma independiente a la variación en la Puntuación Hemorrágica.

```

##          Comparacion Delta_Df      p_value
## 1           Feno → Feno + Geno      10 4.830614e-07
## 2 Feno + Geno → Feno + Geno + VWD 10 1.844483e-10
## 3           Feno → Feno + VWD      10 1.405344e-17
## 4           Geno → Geno + VWD      10 2.235805e-10
## 5           Geno → Feno + Geno      6 3.692063e-01
## 6           VWD → Feno + Geno + VWD 16 1.533726e-01

```

### Interpretación del estudio ANNOVA

Comparación	Resultado	Interpretación
<b>Feno → +Geno</b>	Significativo	Genético aporta información útil
<b>Feno+Geno → +VWD</b>	Muy significativo	VWD es el predictor más fuerte
<b>Feno → +VWD</b>	Muy significativo	VWD mejora al fenotípico
<b>Geno → +VWD</b>	Muy significativo	VWD mejora al genético
<b>Geno → +Feno+Geno</b>	No significativo	Fenotípico no añade nada al genético
<b>VWD →</b>	No significativo	Feno + Geno no añaden nada al VWD
<b>+Feno+Geno+VWD</b>		

### Conclusiones principales (interpretación científica)

1. El **tipo de VWD** es el factor que más explica la Puntuación Hemorrágica (prácticamente domina el modelo).

2. El **cluster genético explica más que el fenotípico**, y este último añade poco.
3. Fenotípico y genético se solapan, especialmente el fenotípico con el genético.
4. La **combinación Feno + Geno no supera a VWD solo**, porque el tipo de VWD captura la gran mayoría de la variación explicable

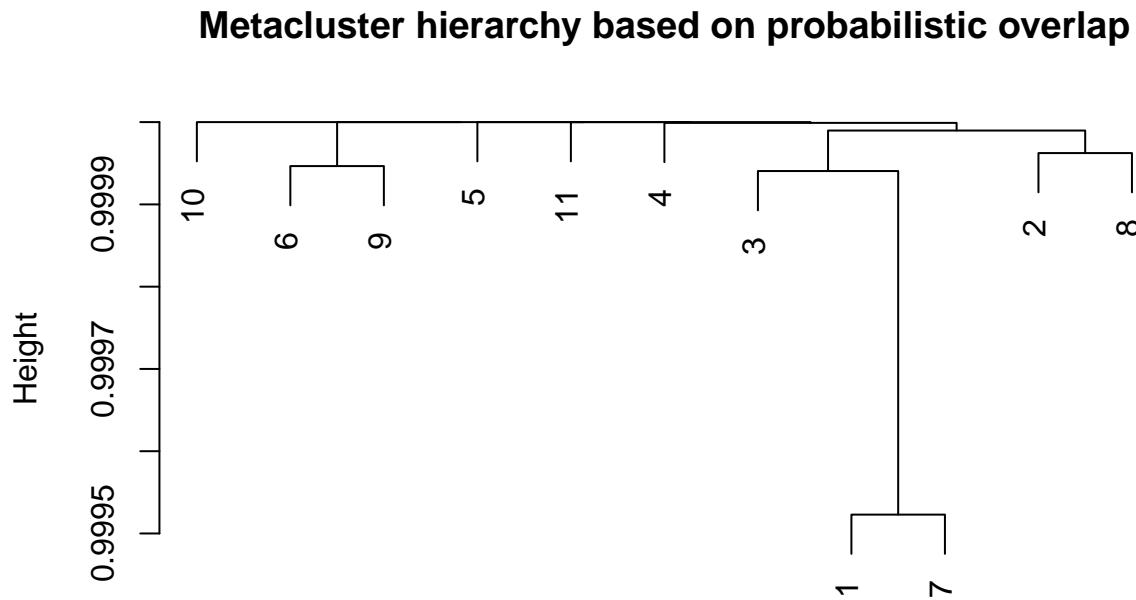
```

z <- geno.mclust$z
G <- ncol(z)

overlap <- t(z) %*% z    # G x G
overlap_norm <- overlap / max(overlap)

d_clusters <- as.dist(1 - overlap_norm)
hc_meta <- hclust(d_clusters, method = "average")
plot(hc_meta, main = "Metacluster hierarchy based on probabilistic overlap")

```



```

d_clusters
hclust (*, "average")

```

```

cluster_max <- apply(z, 1, which.max)
meta_assignment <- meta_clusters[cluster_max]
table(meta_assignment)

## meta_assignment
##   1   2   3   4   5   6   7   8   9   10  11
## 207 120  66  33  28  80  36  38  23  14  15

```

```
heatmap(  
  overlap_norm,  
  symm = TRUE,  
  margins = c(6,6),  
  main = "Probabilistic overlap between Mclust clusters"  
)
```

## Probabilistic overlap between Mclust clusters

