Accurate and robust inference of genetic ancestry from cancer-derived molecular data across genomic platforms

Pascal Belleau[1,2], **Astrid Deschênes**[2,3], **Nyasha Chambwe**[4], **David A. Tuveson**[2,3], **and Alexander Krasnitz**[1,2*]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; [2]Cancer Center, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; [3]Lustgarten Foundation Pancreatic Cancer Research Laboratory, Cold Spring Harbor, New York, USA; [4]Institute of Molecular Medicine, Feinstein Institutes for Medical Research, Northwell Health, Manhasset, New York, USA; [*]**For correspondence:** krasnitz@cshl.edu (AK)

Contents

List of Tables

List of Figures

Supplementary methods: determination of allele fractions

Knowledge of allele fractions (AF) in a cancer-derived profile is a prerequisite for data synthesis (*cf* the Data Synthesis subsection in the Methods and Materials section). We describe a 3-step AF estimation procedure which relies exclusively on the cancer-derived molecular profile, in the absence of a matching cancer-free genotype from the patient, as would be the case for the intended application of our methods. First (step 1), the loss-of-heterozygosity (LOH) regions are delineated. Next (step 2), the regions of allele imbalance where AF differs significantly from 1/2 are identified. Finally (step 3), AF are computed throughout the regions of allele imbalance. These steps are implemented differently, depending on whether the profile originates in the cancer DNA or RNA. We now discuss these steps, in turn for the DNA- and the RNA-derived profiles (Figure SS1).

For the DNA-derived profiles, the LOH regions (step 1) are detected as follows. An LOH region in P must fit into a gap G between any two consecutive HCS positions, where all the observed genotypes are consistent with homozygosity. Any region within G is then considered an LOH region (see Figure S**??** b) if it contains k_1 PHCG positions with $k_1 \geq k_{min}$ and for which the 1KG frequencies F_i, $1 \leq i \leq k_1$ of the alleles observed in the cancer-derived profile P satisfy

$$\log_{10}\left(\prod_{i=1}^{k_1} \frac{F_i^2}{\max\left[F_i^2, (1-F_i)^2, 2F_i(1-F_i)\right]}\right) < \lambda.$$

PHCG positions only are used for this purpose, to reduce correlations due to linkage. The values of k_{min} and λ were chosen so as to maximize, in TCGA OV data set, the overlap between the regions found to be LOH by these criteria and the published LOH regions ASCAT2 files from NCI's Genomic Data Commons (**??**). The latter were determined with full knowledge of the patient's cancer-free genotype. The optimal values were found to be $k_{min} = 3$ and $\lambda = -3$.

Step 2 is based on the notion of an "empty box" (see Figure SS1 b). By this, we mean a contiguous region where the allele fraction of 1/2 is inconsistent with the read counts for the reference and alternative alleles at the HCS positions it contains. An empty box is constructed as follows. First, we consider sliding windows, each encompassing k_2 consecutive HCS positions not separated by an LOH region. A window is called asymmetric if (a) for no less than $k_2 - 1$ of the positions the minor allele count is outside the inner-quartile range (IQR) of the binomial distribution with the minor AF of $f_0 = 1/2$ and (b) satisfy

$$\log_{10}\left(\prod_{i=1}^{k_2} \frac{2P_i}{(1-2P_i)}\right) < \lambda.$$

where P_i = P($X_i \leq$ number of reads covering the minor allele at position i) and, X_i is the binomial distribution with the number of trials equals the coverage at the position i and the probability of success $\rho = 1/2$. In this work, $\lambda = -3$. A polymorphic position is called asymmetric if it belongs to at least one asymmetric window. An empty box is a region with no less than k_2 polymorphic positions, all of which are asymmetric. We used $k_2 = 10$ throughout this work.

At step 3, in the case of DNA, we consider contiguous genome regions of allele asymmetry identified at step 2. Each of these may consist of sub-regions with differing allele fractions. To detect these sub-regions, we "seed" the first sub-region with k_3 HCS positions at the region's boundary and, in this window, estimate the minor allele fraction. We consider the adjacent window W of k_3 HCS positions $k_3 + 1$ through $2k_3$ and apply to it the empty box criteria as described for step 2, with f_0 set to the estimated minor allele fraction of the first window. If the criteria are satisfied, W becomes the seed of the next sub-region, and the process is repeated. Otherwise, HCS position $k_3 + 1$ is added the first sub-region and W is shifted to start at $k_3 + 2$, etc.

In the case of a cancer-derived RNA profile, the expressed allele fractions are, in general, gene specific. Therefore the steps 1 and 2 (condition b), as described above, are performed separately for each gene, assuming the minor allele fraction to be constant throughout the gene. Step 3 is then reduced to an empirical estimate of the minor allele fraction using read counts from all HCS positions within the gene.

Cohort		Sample	
Cohort	Profiling modality	Source	Count
TCGA-OV	WES	Cancer	453
TCGA-OV	WES	Cancer-free	450
TCGA-OV	RNA-seq	Cancer	376
Beat AML	WES	Cancer	343
Beat AML	WES	Cancer-free	532
Beat AML	RNA-seq	Cancer	430
PDAC	WES	Cancer	65
PDAC	WGS	Cancer-free	24
PDAC	RNA-seq	Cancer	40
TCGA-BRCA	WES	Cancer	183
TCGA-BRCA	WES	Cancer-free	183
TCGA-BRCA	RNA-seq	Cancer	183

Table S1. Cancer-derived data used from the TCGA-OV, Beat AML, PDAC and TCGA-BRCA cohorts.

AFR		AMR		EAS		EUR		SAS	
pop	number	pop	number	pop	number	pop	number	pop	number
ACB	95	CLM	94	CDX	96	CEU	96	BEB	85
ASW	50	MXL	62	CHB	106	FIN	105	GIH	103
ESN	96	PEL	84	CHS	101	GBR	96	ITU	101
GWD	109	PUR	104	JPT	105	IBS	107	PJL	91
LWK	88	-	0	KHV	99	TSI	108	STU	99
MSL	82	-	0	-	0	-	0	-	0
YRI	107	-	0	-	0	-	0	-	0
Total	627	-	344	-	507	-	512	-	479
Sampled	210	-	120	-	150	-	150	-	150

Table S2. Super-population and population representation in the 1KG reference data set, with no relatives among the individuals included. For the data synthesis, 30 genotypes were sampled from each 1KG population category, resulting in the sampling from each of the five super-populations as shown.

Table S3. Overall AUROC for 10 TCGA-OV patients, computed using data synthesis.

	P	D	K	Accuracy	95% CI	AUROC	95% CI
WES	P1	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P2	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P3	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P4	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P5	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P6	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P7	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P8	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P9	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P10	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
Panel	P1	4	12	0.9923077	0.986-0.998	0.9944048	0.994-0.995
	P2	4	12	0.9923077	0.986-0.998	0.9944048	0.994-0.995
	P3	4	12	0.9923077	0.986-0.998	0.9944048	0.994-0.995
	P4	4	12	0.9935897	0.988-0.999	0.9954464	0.995-0.996
	P5	4	12	0.9935897	0.988-0.999	0.9952381	0.995-0.995
	P6	4	12	0.9923077	0.986-0.998	0.9944048	0.994-0.995
	P7	4	12	0.9948718	0.99-1	0.9962798	0.996-0.996
	P8	4	12	0.9935897	0.988-0.999	0.9952381	0.995-0.995
	P9	4	12	0.9910256	0.984-0.998	0.9935714	0.993-0.994
	P10	4	12	0.9923077	0.986-0.998	0.9944048	0.994-0.995
RNA-seq	P1	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998
	P2	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998
	P3	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998
	P4	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998
	P5	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998
	P6	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998
	P7	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998
	P8	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998
	P9	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998
	P10	7	12	0.9974359	0.994-1	0.9983631	0.998-0.998

Table S4. Overall AUROC for 10 Beat AML patients, computed using data synthesis.

	P	D	K	Accuracy	95% CI	AUROC	95% CI
WES	P1	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P2	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P3	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P4	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P5	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P6	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P7	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P8	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P9	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P10	5	13	0.9974359	0.994-1	0.9983631	0.998-0.998
Panel	P1	4	13	0.9961538	0.992-1	0.9973214	0.997-0.997
	P2	4	13	0.9935897	0.988-0.999	0.9954464	0.995-0.996
	P3	4	13	0.9961538	0.992-1	0.9973214	0.997-0.997
	P4	4	13	0.9961538	0.992-1	0.9973214	0.997-0.997
	P5	4	13	0.9961538	0.992-1	0.9973214	0.997-0.997
	P6	4	13	0.9935897	0.988-0.999	0.9956845	0.996-0.996
	P7	4	13	0.9948718	0.99-1	0.9962798	0.996-0.996
	P8	4	13	0.9948718	0.99-1	0.9967262	0.997-0.997
	P9	4	13	0.9935897	0.988-0.999	0.9956845	0.996-0.996
	P10	4	13	0.9935897	0.988-0.999	0.9956845	0.996-0.996
RNA-seq	P1	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P2	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P3	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P4	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P5	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P6	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P7	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P8	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P9	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P10	4	13	0.9974359	0.994-1	0.9983631	0.998-0.998

Table S5. Overall AUROC for 10 PDAC patients, computed using data synthesis.

	P	D	K	Accuracy	95% CI	AUROC	95% CI
WES	P1	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P2	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P3	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P4	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P5	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P6	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P7	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P8	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P9	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
	P10	8	13	0.9974359	0.994-1	0.9983631	0.998-0.998
Panel	P1	6	5	0.9923077	0.986-0.998	0.9950595	0.995-0.995
	P2	6	5	0.9948718	0.99-1	0.9964881	0.996-0.997
	P3	6	5	0.9935897	0.988-0.999	0.9958929	0.996-0.996
	P4	6	5	0.9948718	0.99-1	0.9964881	0.996-0.997
	P5	6	5	0.9961538	0.992-1	0.9973214	0.997-0.997
	P6	6	5	0.9923077	0.986-0.998	0.9950595	0.995-0.995
	P7	6	5	0.9961538	0.992-1	0.9973214	0.997-0.997
	P8	6	5	0.9935897	0.988-0.999	0.9958929	0.996-0.996
	P9	6	5	0.9961538	0.992-1	0.9975298	0.997-0.998
	P10	6	5	0.9948718	0.99-1	0.9964881	0.996-0.997
RNA-seq	P1	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998
	P2	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998
	P3	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998
	P4	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998
	P5	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998
	P6	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998
	P7	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998
	P8	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998
	P9	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998
	P10	5	11	0.9974359	0.994-1	0.9983631	0.998-0.998

Table S6. Overall AUROC for 10 TCGA-BRCA patients, computed using data synthesis.

	P	D	K	Accuracy	95% CI	AUROC	95% CI
WES	P1	4	9	0.9958718	0.985-1	0.9962798	0.996-0.996
	P2	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P3	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P4	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P5	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P6	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P7	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P8	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P9	4	8	0.9948718	0.985-1	0.9962798	0.996-0.996
	P10	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
Panel	P1	4	10	0.9910256	0.977-1	0.9931548	0.993-0.993
	P2	4	11	0.9910256	0.977-1	0.9931548	0.993-0.993
	P3	4	6	0.9910256	0.977-1	0.9931548	0.993-0.993
	P4	3	9	0.9871795	0.971-1	0.9910714	0.991-0.991
	P5	4	3	0.9910256	0.977-1	0.9931548	0.993-0.993
	P6	4	6	0.9935897	0.982-1	0.9952381	0.995-0.995
	P7	4	9	0.9923077	0.98-1	0.9941964	0.994-0.994
	P8	4	3	0.9884615	0.973-1	0.9919643	0.992-0.992
	P9	4	7	0.9910256	0.977-1	0.9933631	0.993-0.993
	P10	4	5	0.9897436	0.975-1	0.9923214	0.992-0.992
RNA-seq	P1	4	11	0.9948718	0.985-1	0.9962798	0.996-0.996
	P2	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P3	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P4	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P5	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P6	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P7	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P8	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P9	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996
	P10	4	9	0.9948718	0.985-1	0.9962798	0.996-0.996

Source and modality	D	K	Accuracy	95% CI	AUROC	95% CI
TCGA-OV Cancer-free WES	5	13	0.986	0.975-0.997	0.994	0.994-0.995
TCGA-OV Cancer WES	5	13	0.990	0.980-1	0.995	0.994-0.995
TCGA-OV Cancer panel	4	12	0.991	0.982-1	0.995	0.994-0.995
TCGA-OV Cancer RNA-seq	7	12	0.989	0.978-1	0.978	0.976-0.981
TCGA-BRCA Cancer-free WES	9	4	0.932	0.884-0.964	0.978	0.977-0.979
TCGA-BRCA Cancer WES	9	4	0.931	0.882-0.964	0.978	0.977-0.979
TCGA-BRCA Cancer panel	9	4	0.931	0.882-0.964	0.978	0.977-0.979
TCGA-BRCA Cancer RNA-seq	9	4	0.931	0.883-0.964	0.978	0.977-0.979

Table S7. Cohort-wide performance measures for super-population inference from TCGA-OV and TCGA-BRCA molecular data, with the C5 (***Carrot-Zhang et al., 2020***) ancestry calls as the ground truth.

Super-population population		Source and modality			
		Cancer-free WES	Cancer WES	Cancer panel	Cancer RNA-seq
EAS					
	AUROC	1	1	1	1
	95% CI	1-1	1-1	1-1	1-1
EUR					
	AUROC	0.913	0.912	0.912	0.912
	95% CI	0.868-0.958	0.866-0.957	0.866-0.957	0.866-0.957
AFR					
	AUROC	1	1	1	1
	95% CI	1-1	1-1	1-1	1-1
AMR					
	AUROC	0.965	0.964	0.967	0.967
	95% CI	0.946-0.984	0.945-0.984	0.949-0.986	0.949-0.986
SAS					
	AUROC	1	1	0.997	0.997
	95% CI	1-1	1-1	0.991-1	0.991-1

Table S8. Cohort-wide performance, quantified as AUROC for super-population inference from TCGA-BRCA molecular data, with the C5 (***Carrot-Zhang et al., 2020***) ancestry calls as the ground truth.

(a) WES Normal

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
C5	EAS	10	0	0	0	0
	EUR	0	383	0	5	0
	AFR	0	0	30	1	0
	AMR	0	0	0	1	0
	SAS	0	0	0	0	6

(b) WES

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
C5	EAS	11	0	0	0	0
	EUR	0	386	0	4	0
	AFR	0	0	29	1	0
	AMR	0	0	0	1	0
	SAS	0	0	0	0	6

(c) RNA-seq

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
C5	EAS	7	1	0	0	0
	EUR	0	325	0	2	0
	AFR	0	0	23	1	0
	AMR	0	1	0	1	0
	SAS	0	0	0	0	4

(d) Panel

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
C5	EAS	11	0	0	0	0
	EUR	0	387	0	3	0
	AFR	0	0	29	1	0
	AMR	0	0	0	1	0
	SAS	0	0	0	0	6

Table S9. Confusion matrices for super-population calls from TCGA-OV cancer-derived data, in comparison to the C5 calls (***Carrot-Zhang et al., 2020***)

(a) WES Normal

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
C5	EAS	47	0	0	0	0
	EUR	0	56	0	12	0
	AFR	0	0	50	0	0
	AMR	0	0	0	4	0
	SAS	0	0	0	0	4

(b) WES Cancer

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
C5	EAS	47	0	0	0	0
	EUR	0	56	0	12	0
	AFR	0	0	50	0	0
	AMR	0	0	0	4	0
	SAS	0	0	0	0	4

(c) RNA-seq

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
C5	EAS	47	0	0	0	0
	EUR	0	56	0	11	1
	AFR	0	0	50	0	0
	AMR	0	0	0	4	0
	SAS	0	0	0	0	4

(d) Panel

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
C5	EAS	47	0	0	0	0
	EUR	0	56	0	11	1
	AFR	0	0	50	0	0
	AMR	0	0	0	4	0
	SAS	0	0	0	0	4

Table S10. Confusion matrices for super-population calls from TCGA-BRCA cancer-derived data, in comparison to the C5 calls (***Carrot-Zhang et al., 2020***)

Super-population population		Source and modality					
		TCGA-BRCA WES	TCGA-BRCA Panel	TCGA-BRCA RNA-seq	Aggregate WES	Aggregate Panel	Aggregate RNA-seq
EAS							
	AUROC	1	1	1	1	1	0.999
	95% CI	1-1	1-1	1-1	1-1	1-1	0.998-1
EUR							
	AUROC	1	1	1	0.994	0.987	0.996
	95% CI	1-1	1-1	1-1	0.989-0.999	0.977-0.996	0.991-1
AFR							
	AUROC	1	1	1	1	0.995	1
	95% CI	1-1	1-1	1-1	1-1	0.985-1	1-1
AMR							
	AUROC	1	0.980	0.980	0.990	0.954	0.974
	95% CI	1-1	0.941-1	0.941-1	0.976-1	0.922-0.987	0.947-1
SAS							
	AUROC	1	0.997	0.997	1	0.999	0.999
	95% CI	1-1	0.992-1	0.992-1	1-1	0.998-1	0.998-1

Table S11. Cohort-wide performance, quantified as AUROC for super-population inference from TCGA-BRCA molecular data, with the cancer-free WES ancestry calls as the ground truth.

(a) WES

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WGS	EAS	1	0	0	0	0
	EUR	0	15	0	0	0
	AFR	0	0	2	0	0
	AMR	0	0	0	2	0
	SAS	0	0	0	0	1

(b) Panel

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WGS	EAS	1	0	0	0	0
	EUR	0	15	0	0	0
	AFR	0	0	2	0	0
	AMR	0	1	0	1	0
	SAS	0	0	0	0	1

(c) RNA-seq

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WGS	EAS	1	0	0	0	0
	EUR	0	13	0	0	0
	AFR	0	0	2	0	0
	AMR	0	0	0	2	0
	SAS	0	0	0	0	1

Table S12. Confusion matrices for super-population calls from PDAC cancer-derived data, in comparison to those from cancer-free WGS.

(a) WES

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	10	0	0	0	0
	EUR	0	378	0	0	0
	AFR	0	0	29	0	0
	AMR	0	1	0	16	0
	SAS	0	0	0	0	7
	UNK	0	2	0	0	0

(b) Panel

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	10	0	0	0	0
	EUR	0	376	0	2	0
	AFR	0	0	28	1	0
	AMR	0	4	0	13	0
	SAS	0	0	0	0	7
	UNK	0	2	0	0	0

(c) RNA-seq

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	4	0	0	0	0
	EUR	0	242	0	0	0
	AFR	0	0	21	0	0
	AMR	1	1	0	9	0
	SAS	0	0	0	0	4
	UNK	0	1	0	0	0

Table S13. Confusion matrices for super-population calls from TCGA-OV cancer-derived data, in comparison to those from cancer-free WES.

(a) WES

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	11	0	0	0	0
	EUR	0	283	0	6	0
	AFR	0	0	14	0	0
	AMR	0	0	0	27	0
	SAS	0	0	0	0	2
	UNK	0	0	0	0	0

(b) Panel

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	11	0	0	0	0
	EUR	0	286	0	3	0
	AFR	0	0	14	0	0
	AMR	0	0	0	27	0
	SAS	0	0	0	0	2
	UNK	0	0	0	0	0

(c) RNA-seq

		Inferred				
	pop	EAS	EUR	AFR	AMR	SAS
Cancer-free WES	EAS	10	0	0	0	0
	EUR	0	210	0	2	0
	AFR	0	0	9	0	0
	AMR	0	0	0	24	0
	SAS	0	0	0	0	1
	UNK	0	0	0	0	0

Table S14. Confusion matrices for super-population calls from Beat AML cancer-derived data, in comparison to those from cancer-free WES.

Table S15. Super-population specific AUROC for 10 TCGA-OV patients, computed using data synthesis.

	P	EAS	95% CI	EUR	95% CI	AFR	95% CI	AMR	95% CI	SAS	95% CI
WES	P1	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P2	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P3	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P4	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P5	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P6	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P7	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P8	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P9	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P10	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
Panel	P1	1	1-1	0.994	0.987-1	0.997	0.992-1	0.982	0.966-0.998	1	1-1
	P2	1	1-1	0.994	0.987-1	0.997	0.992-1	0.982	0.966-0.998	1	1-1
	P3	1	1-1	0.994	0.987-1	0.997	0.992-1	0.982	0.966-0.998	1	1-1
	P4	1	1-1	0.995	0.988-1	0.997	0.992-1	0.986	0.972-1	1	1-1
	P5	1	1-1	0.998	0.995-1	0.997	0.992-1	0.983	0.966-0.999	1	1-1
	P6	1	1-1	0.994	0.987-1	0.997	0.992-1	0.982	0.966-0.998	1	1-1
	P7	1	1-1	0.998	0.996-1	0.997	0.992-1	0.987	0.973-1	1	1-1
	P8	1	1-1	0.998	0.995-1	0.997	0.992-1	0.983	0.966-0.999	1	1-1
	P9	1	1-1	0.991	0.981-1	0.997	0.992-1	0.981	0.965-0.997	1	1-1
	P10	1	1-1	0.994	0.987-1	0.997	0.992-1	0.982	0.966-0.998	1	1-1
RNA-seq	P1	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P2	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P3	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P4	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P5	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P6	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P7	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P8	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P9	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P10	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1

Table S16. Super-population specific AUROC for 10 Beat AML patients, computed using data synthesis.

	P	EAS	95% CI	EUR	95% CI	AFR	95% CI	AMR	95% CI	SAS	95% CI
WES	P1	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P2	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P3	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P4	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P5	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P6	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P7	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P8	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P9	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P10	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
Panel	P1	1	1-1	0.999	0.998-1	0.997	0.992-1	0.991	0.979-1	1	1-1
	P2	1	1-1	0.995	0.988-1	0.997	0.992-1	0.986	0.972-1	1	1-1
	P3	1	1-1	0.999	0.998-1	0.997	0.992-1	0.991	0.979-1	1	1-1
	P4	1	1-1	0.999	0.998-1	0.997	0.992-1	0.991	0.979-1	1	1-1
	P5	1	1-1	0.999	0.998-1	0.997	0.992-1	0.991	0.979-1	1	1-1
	P6	1	1-1	0.998	0.996-1	0.994	0.988-1	0.986	0.972-1	1	1-1
	P7	1	1-1	0.998	0.996-1	0.997	0.992-1	0.987	0.973-1	1	1-1
	P8	1	1-1	0.999	0.998-1	0.994	0.988-1	0.990	0.978-1	1	1-1
	P9	1	1-1	0.998	0.996-1	0.994	0.988-1	0.986	0.972-1	1	1-1
	P10	1	1-1	0.998	0.996-1	0.994	0.988-1	0.986	0.972-1	1	1-1
RNA-seq	P1	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P2	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P3	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P4	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P5	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P6	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P7	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P8	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P9	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P10	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1

Table S17. Super-population specific AUROC for 10 PDAC patients, computed using data synthesis.

	P	EAS	95% CI	EUR	95% CI	AFR	95% CI	AMR	95% CI	SAS	95% CI
WES	P1	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P2	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P3	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P4	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P5	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P6	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P7	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P8	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P9	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P10	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
Panel	P1	1	1-1	0.993	0.983-1	0.994	0.988-1	0.989	0.977-1	1	1-1
	P2	1	1-1	0.996	0.989-1	0.997	0.992-1	0.990	0.978-1	1	1-1
	P3	1	1-1	0.996	0.989-1	0.994	0.988-1	0.989	0.978-1	1	1-1
	P4	1	1-1	0.996	0.989-1	0.997	0.992-1	0.990	0.978-1	1	1-1
	P5	1	1-1	0.999	0.998-1	0.997	0.992-1	0.991	0.979-1	1	1-1
	P6	1	1-1	0.993	0.983-1	0.994	0.988-1	0.989	0.977-1	1	1-1
	P7	1	1-1	0.999	0.998-1	0.997	0.992-1	0.991	0.979-1	1	1-1
	P8	1	1-1	0.996	0.989-1	0.994	0.988-1	0.989	0.978-1	1	1-1
	P9	1	1-1	0.997	0.99-1	0.997	0.992-1	0.994	0.986-1	1	1-1
	P10	1	1-1	0.996	0.989-1	0.997	0.992-1	0.990	0.978-1	1	1-1
RNA-seq	P1	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P2	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P3	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P4	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P5	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P6	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P7	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P8	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P9	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1
	P10	1	1-1	1.000	1-1	0.997	0.992-1	0.995	0.987-1	1	1-1

Table S18. Super-population specific AUROC for 10 TCGA-BRCA patients, computed using data synthesis.

	P	EAS	95% CI	EUR	95% CI	AFR	95% CI	AMR	95% CI	SAS	95% CI
WES	P1	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P2	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P3	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P4	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P5	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P6	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P7	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P8	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P9	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P10	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
Panel	P1	1	1-1	0.998	0.995-1	0.995	0.989-1	0.974	0.955-0.994	1	1-1
	P2	1	1-1	0.998	0.995-1	0.995	0.989-1	0.974	0.955-0.994	1	1-1
	P3	1	1-1	0.998	0.995-1	0.995	0.989-1	0.974	0.955-0.994	1	1-1
	P4	1	1-1	0.983	0.968-0.997	0.995	0.989-1	0.979	0.962-0.995	1	1-1
	P5	1	1-1	0.998	0.995-1	0.995	0.989-1	0.974	0.955-0.994	1	1-1
	P6	1	1-1	0.999	0.998-1	0.995	0.989-1	0.983	0.966-0.999	1	1-1
	P7	1	1-1	0.998	0.995-1	0.995	0.989-1	0.978	0.96-0.996	1	1-1
	P8	1	1-1	0.998	0.995-1	0.990	0.982-0.999	0.973	0.953-0.992	1	1-1
	P9	1	1-1	0.995	0.988-1	0.995	0.989-1	0.978	0.96-0.996	1	1-1
	P10	1	1-1	0.994	0.987-1	0.995	0.989-1	0.973	0.954-0.993	1	1-1
RNA-seq	P1	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P2	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P3	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P4	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P5	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P6	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P7	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P8	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P9	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1
	P10	1	1-1	1.000	1-1	0.995	0.989-1	0.987	0.973-1	1	1-1

Estimate the allelic ratio for each gene using phasing and haplotype blocks

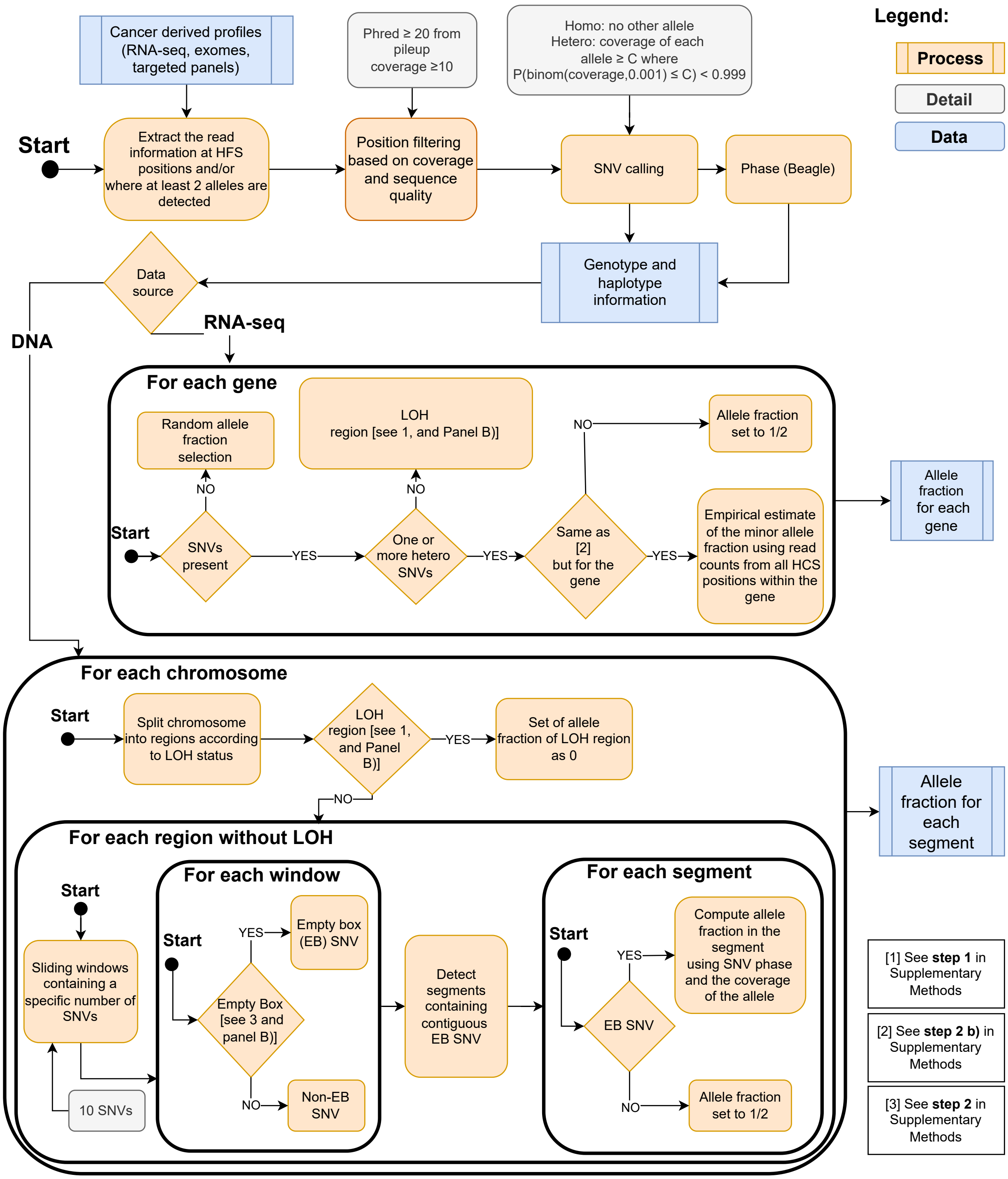

Figure S1. A) Overview of the procedure for allele fraction estimation in genes and segments.

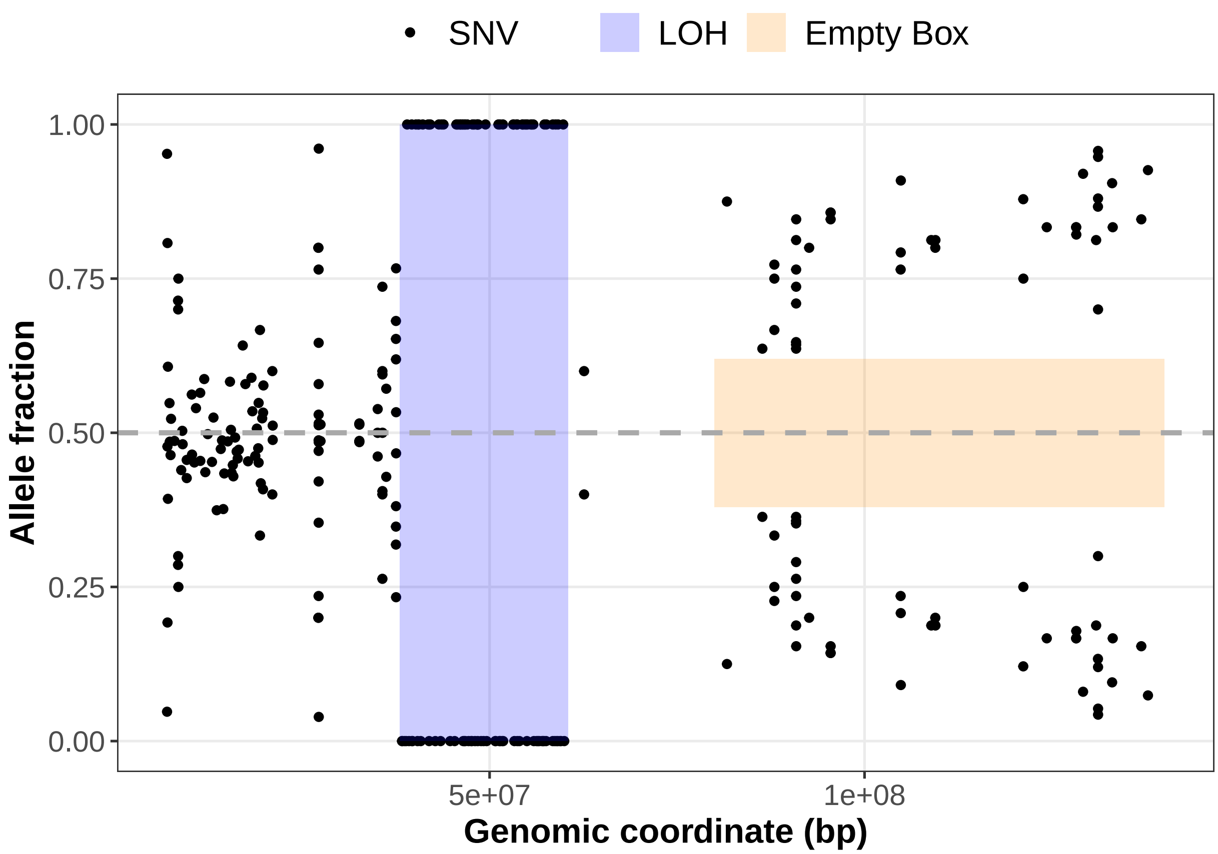

Figure S1 B) Illustration of the key concepts used in estimation of the allele fractions. Each data point represents an observed variant allele fraction at the corresponding genomic coordinate.

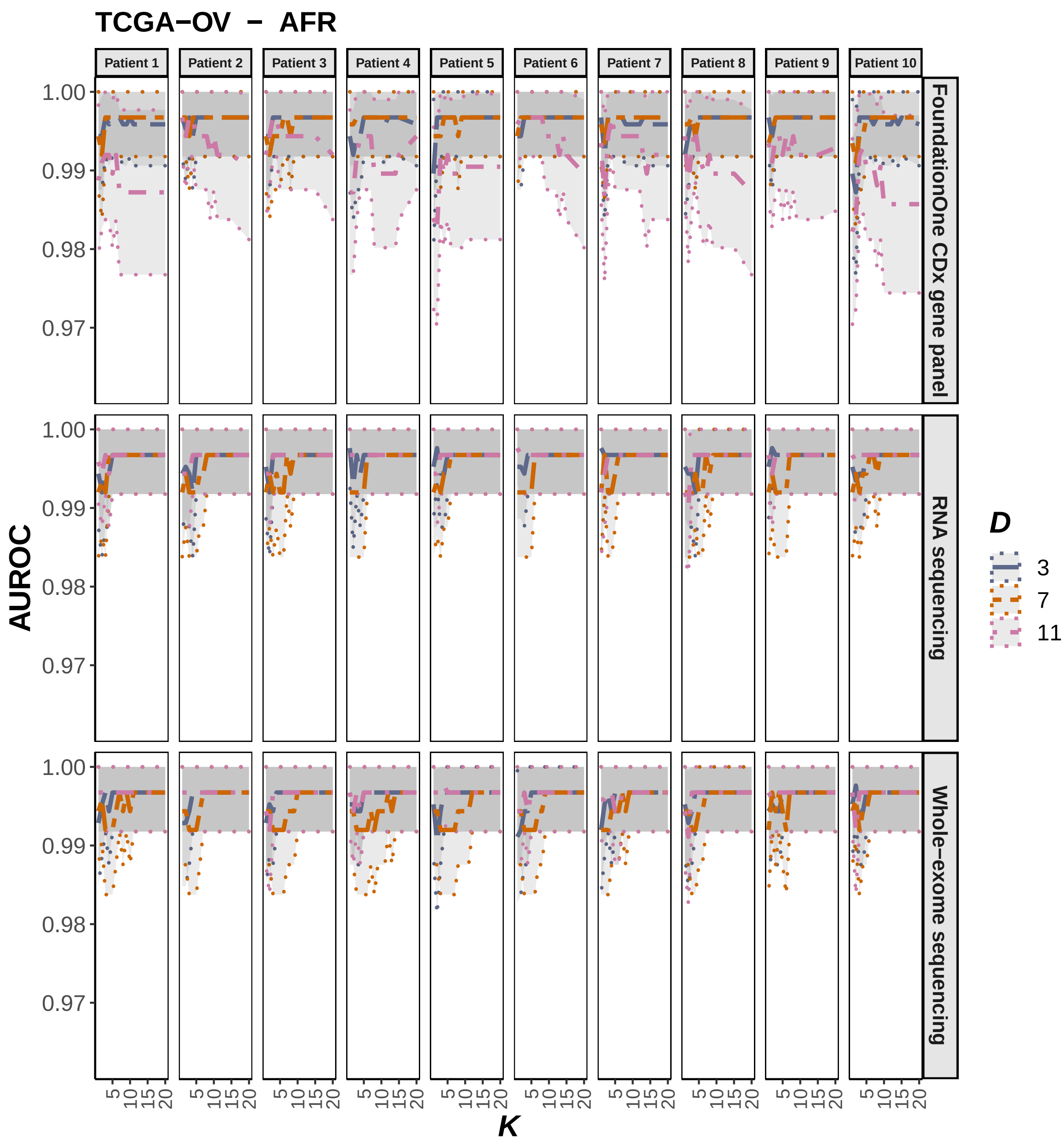

Figure S2. A) Dependence of AFR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-OV patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

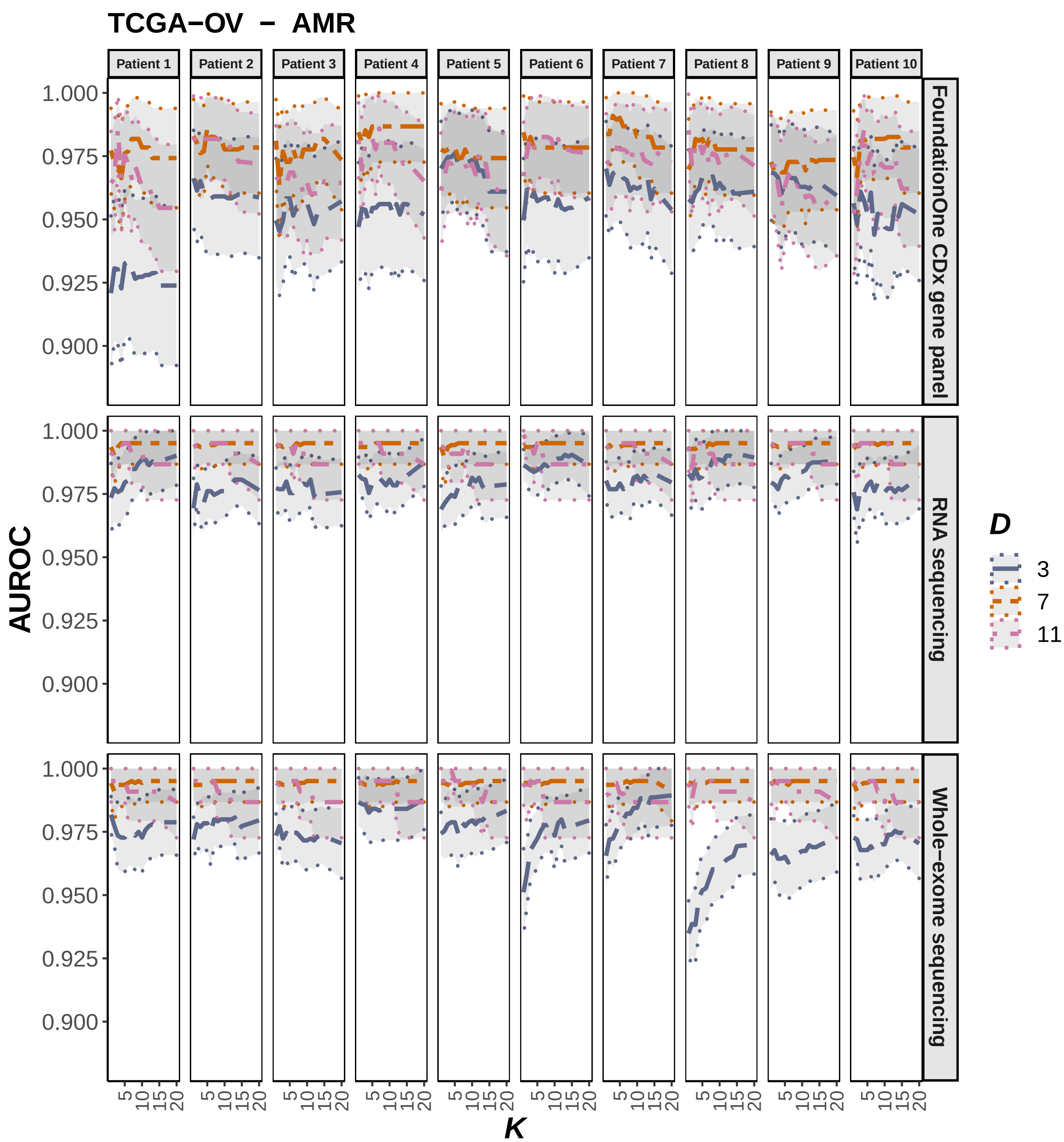

Figure S2. B) Dependence of AMR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-OV patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

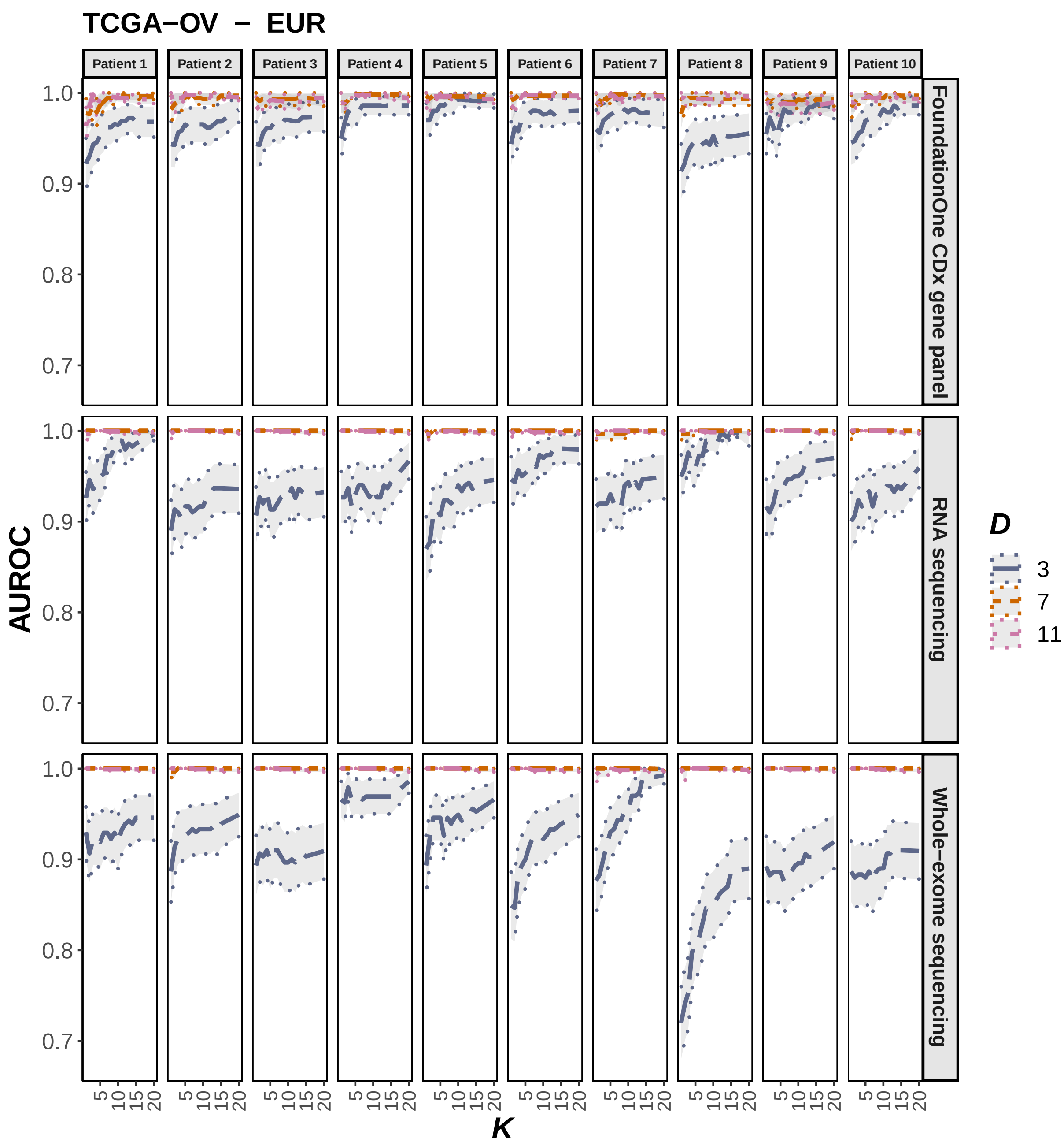

Figure S2. C) Dependence of EUR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-OV patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

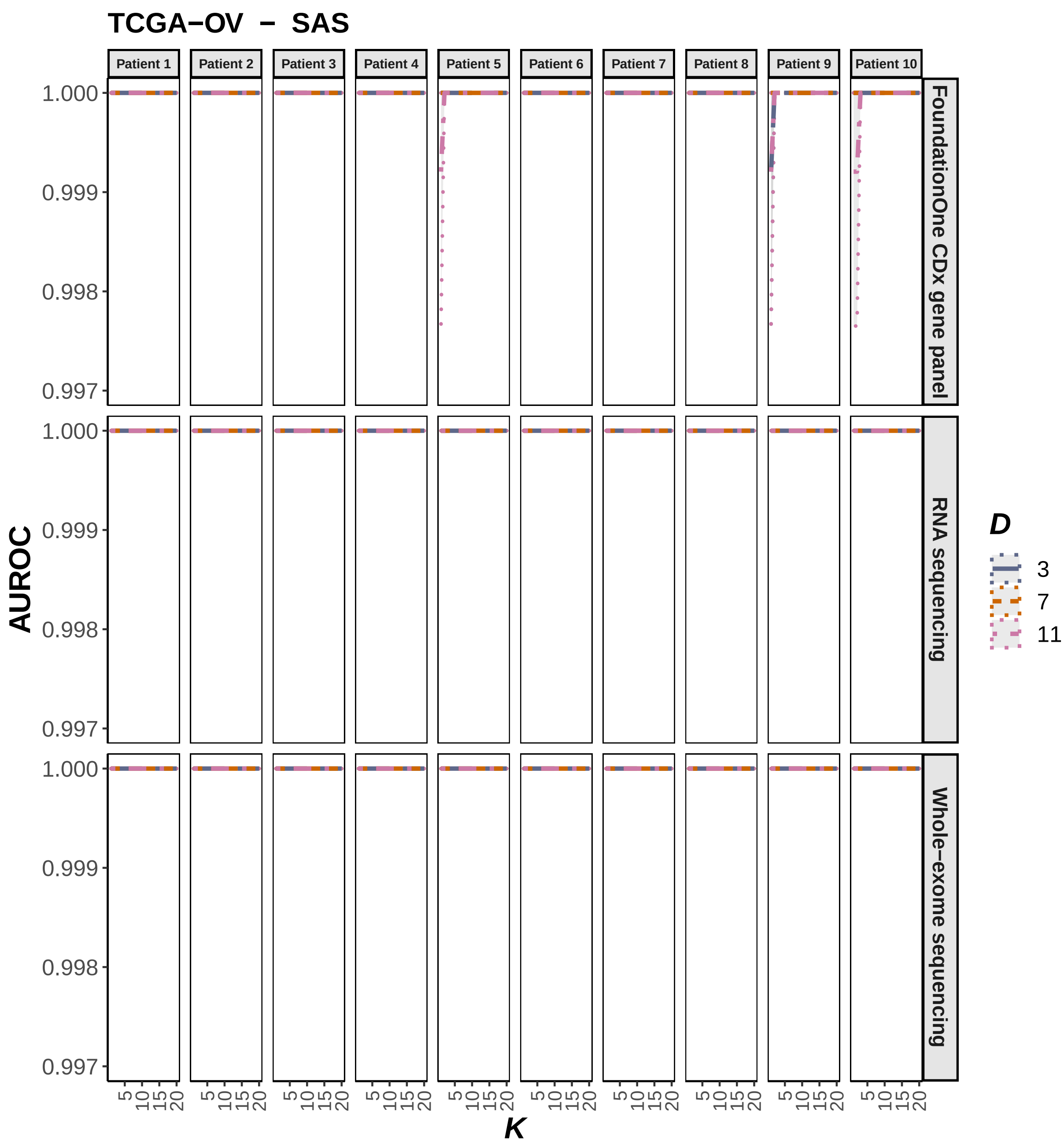

Figure S2. D) Dependence of SAS-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-OV patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

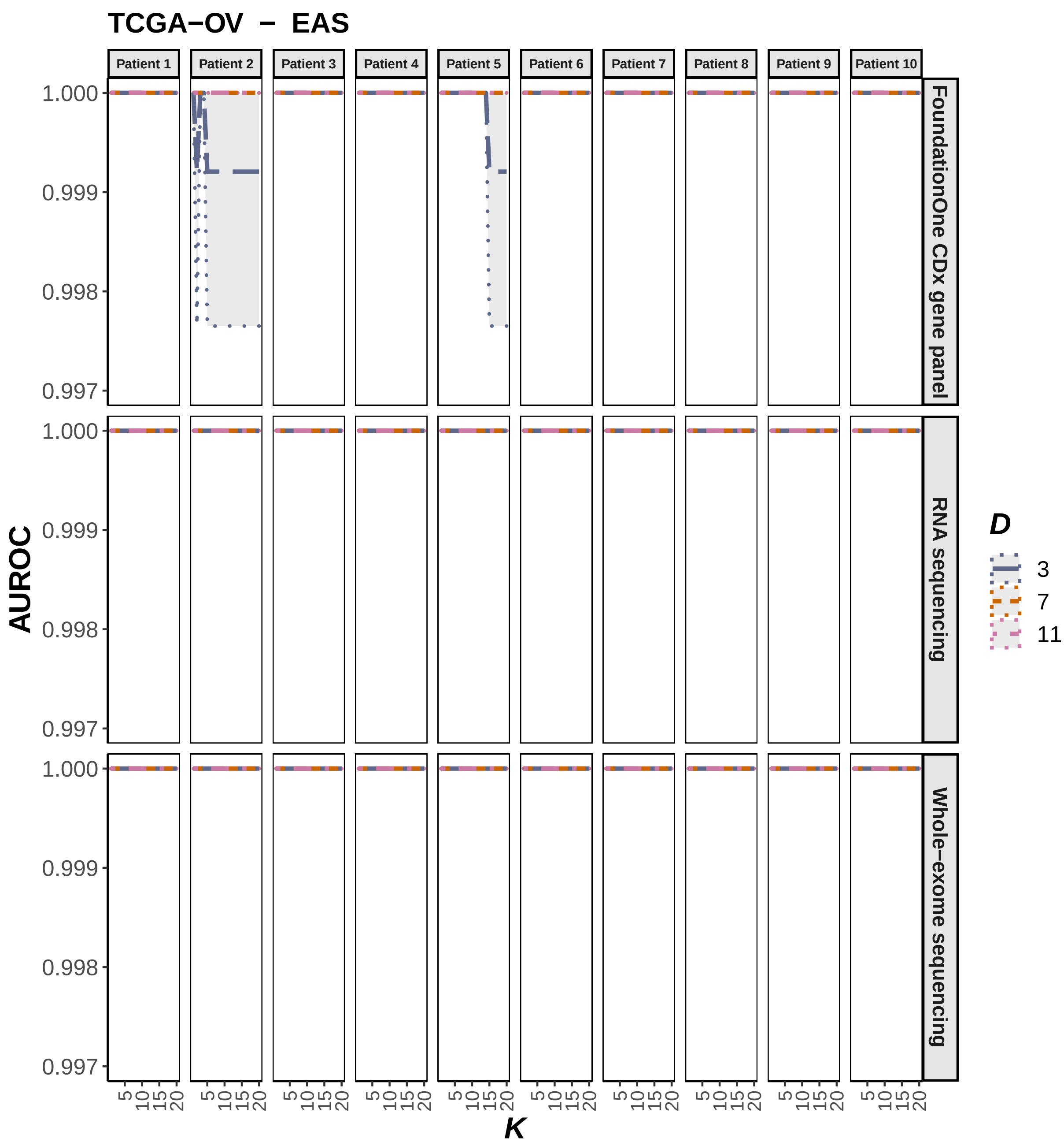

Figure S2. E) Dependence of EAS-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-OV patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

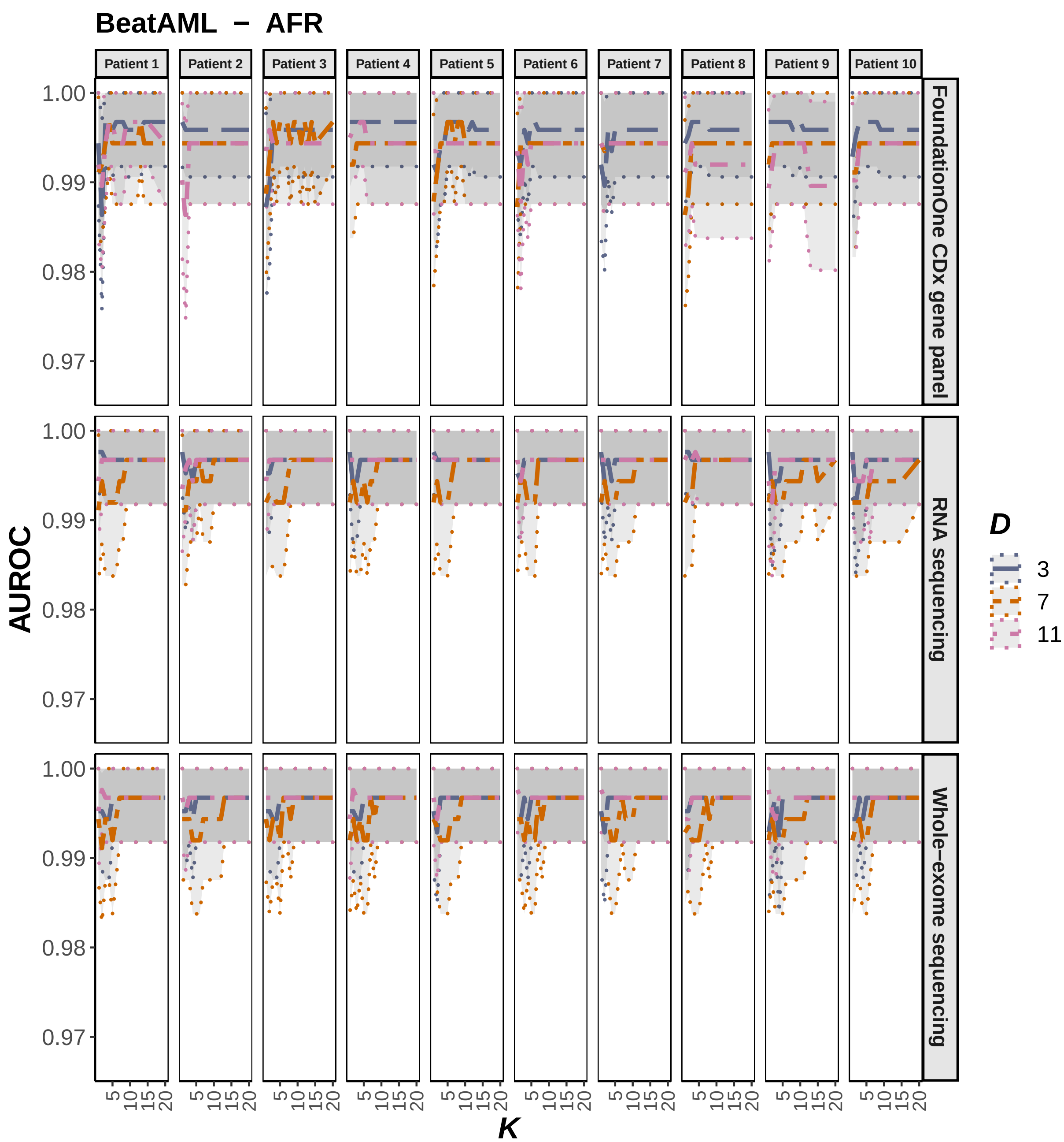

Figure S2. F) Dependence of AFR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 Beat AML patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

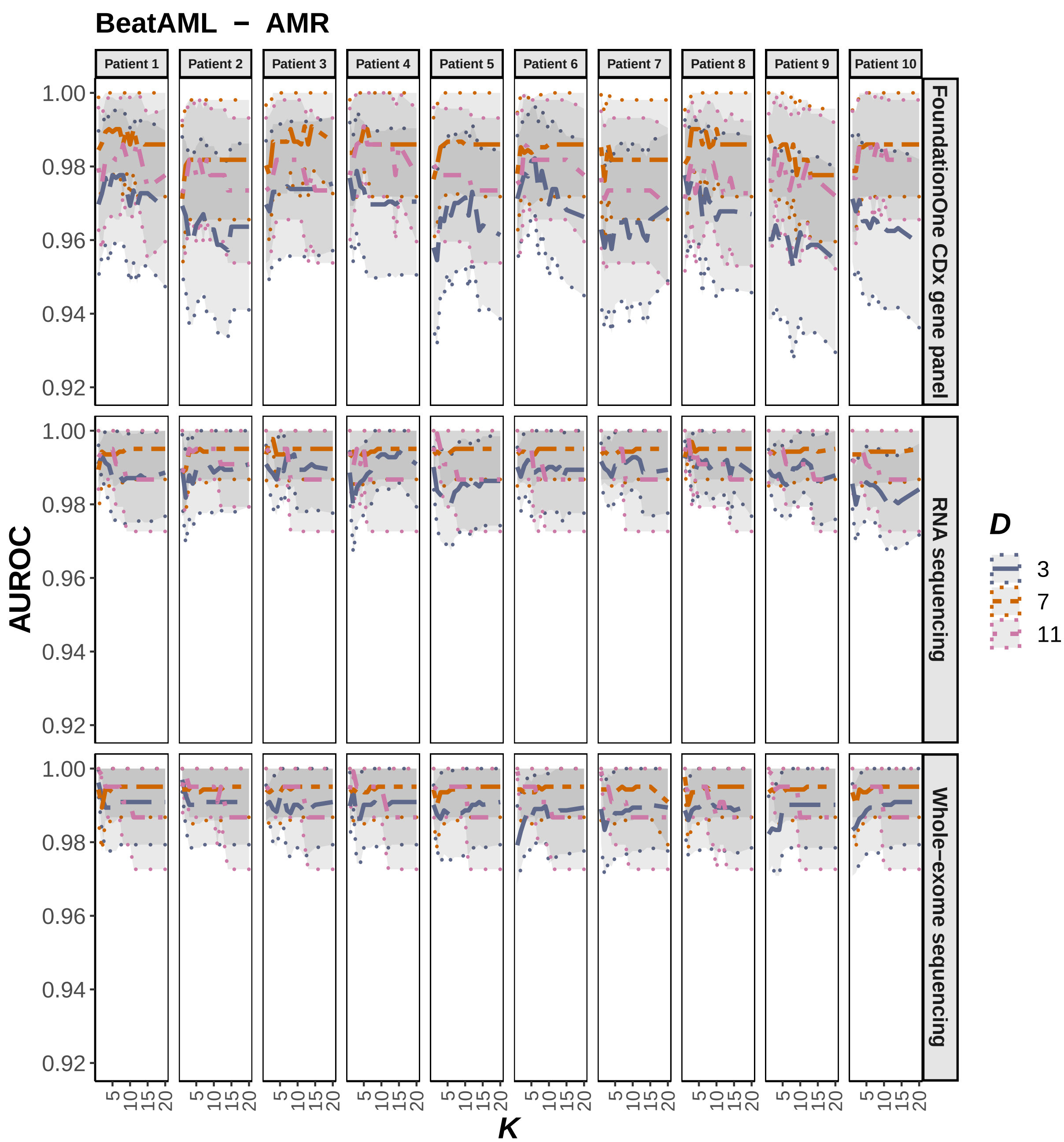

Figure S2. G) Dependence of AMR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 Beat AML patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

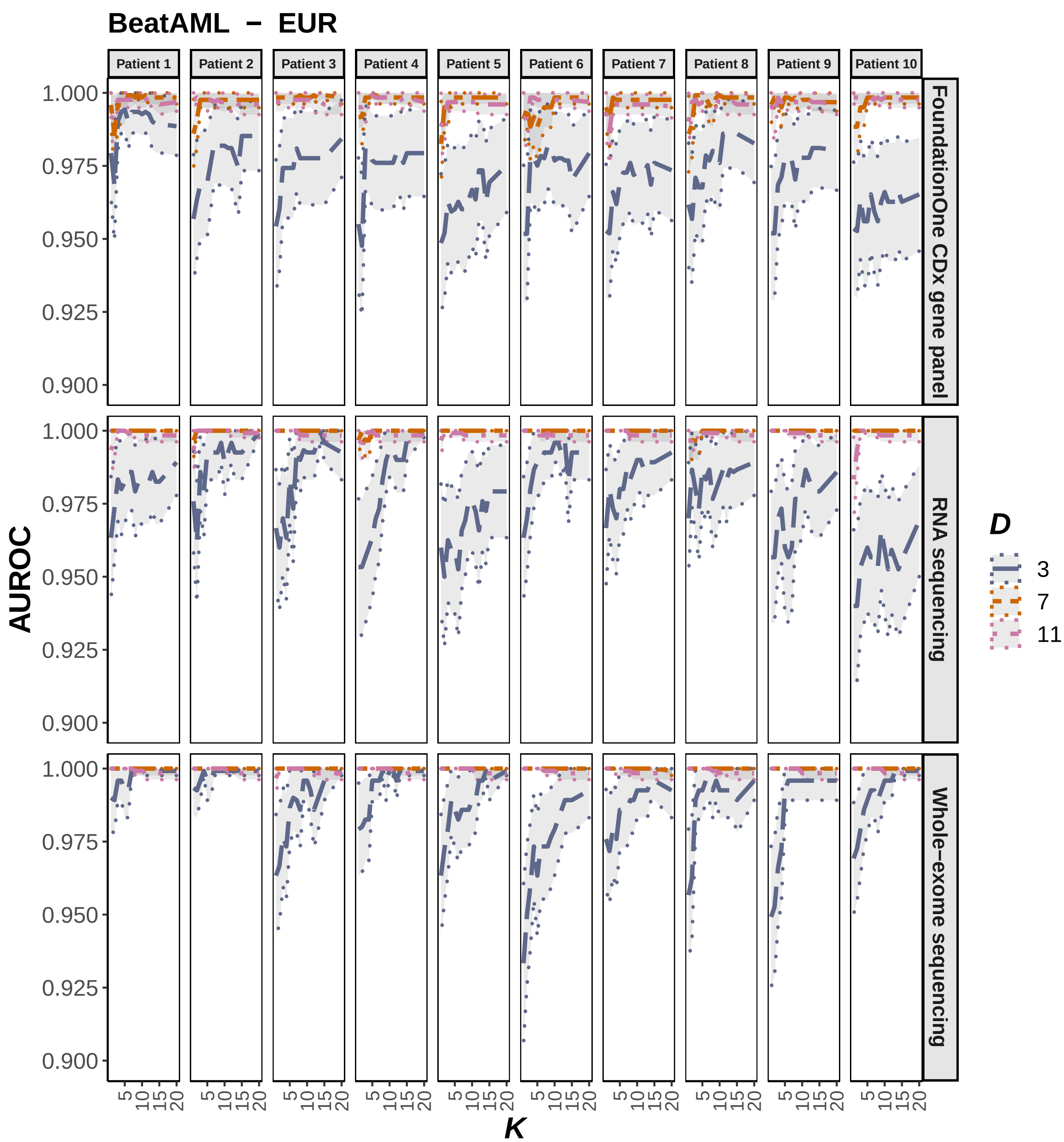

Figure S2. H) Dependence of EUR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 Beat AML patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

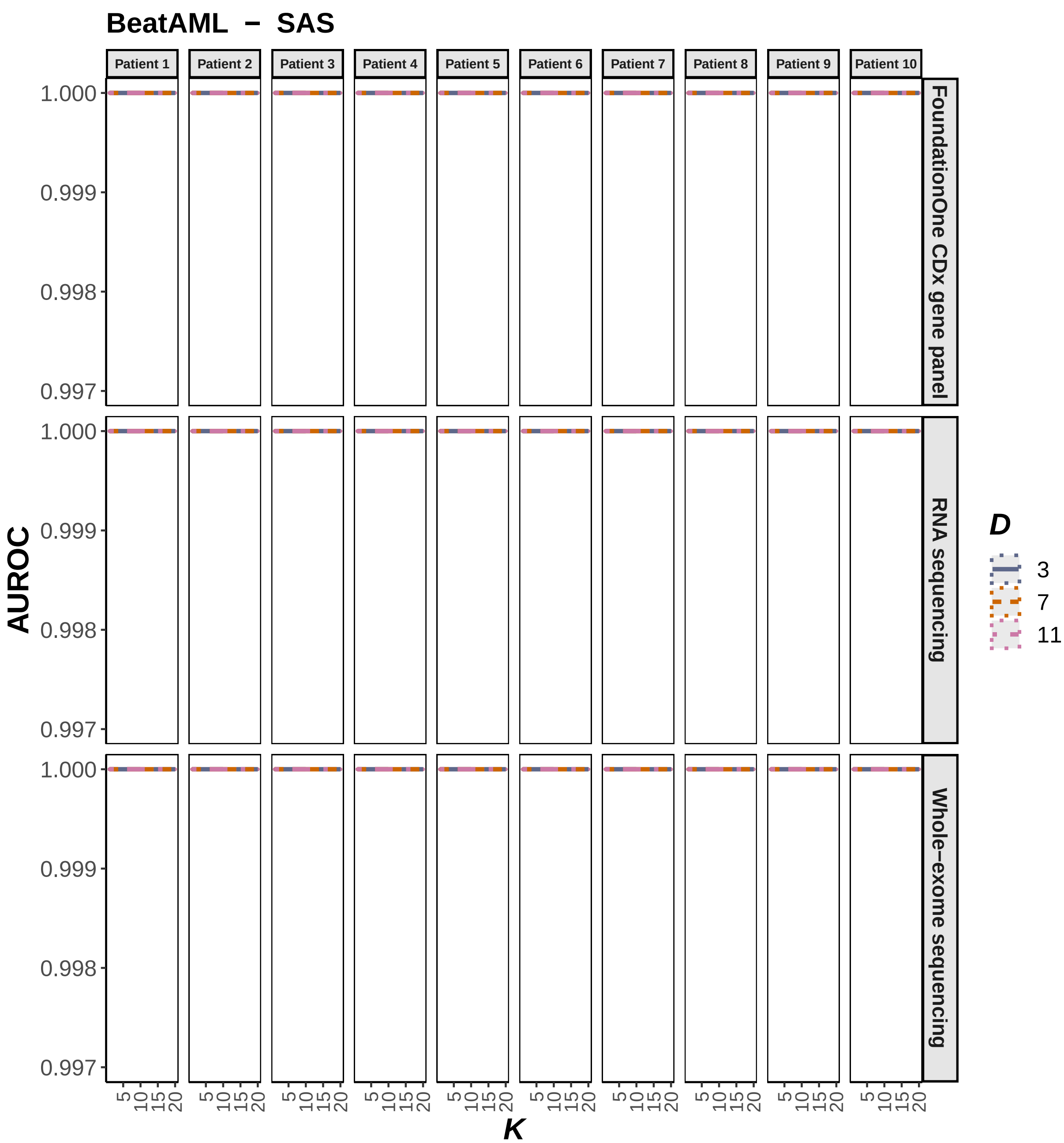

Figure S2. I) Dependence of SAS-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 Beat AML patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

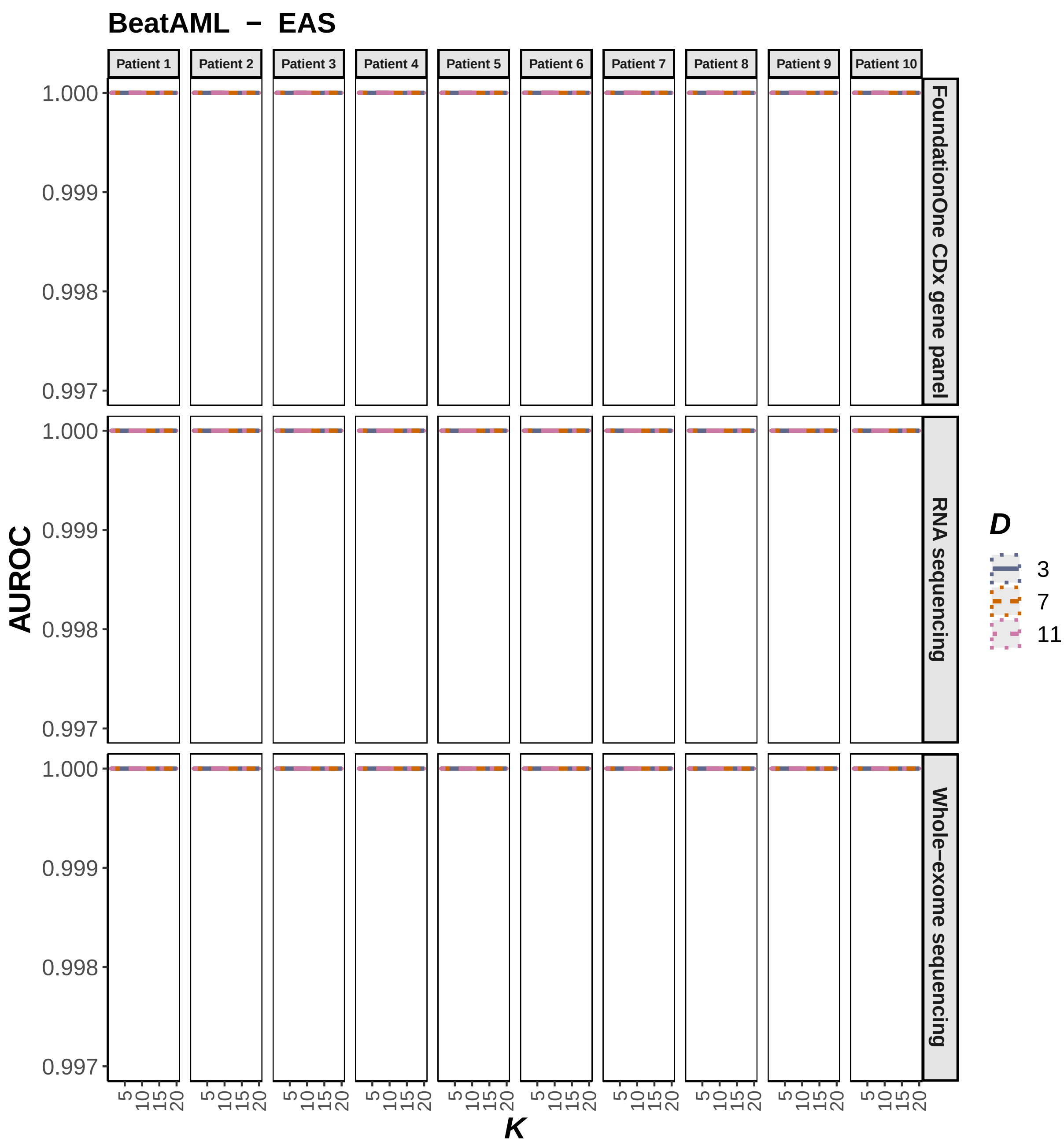

Figure S2. J) Dependence of EAS-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 Beat AML patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

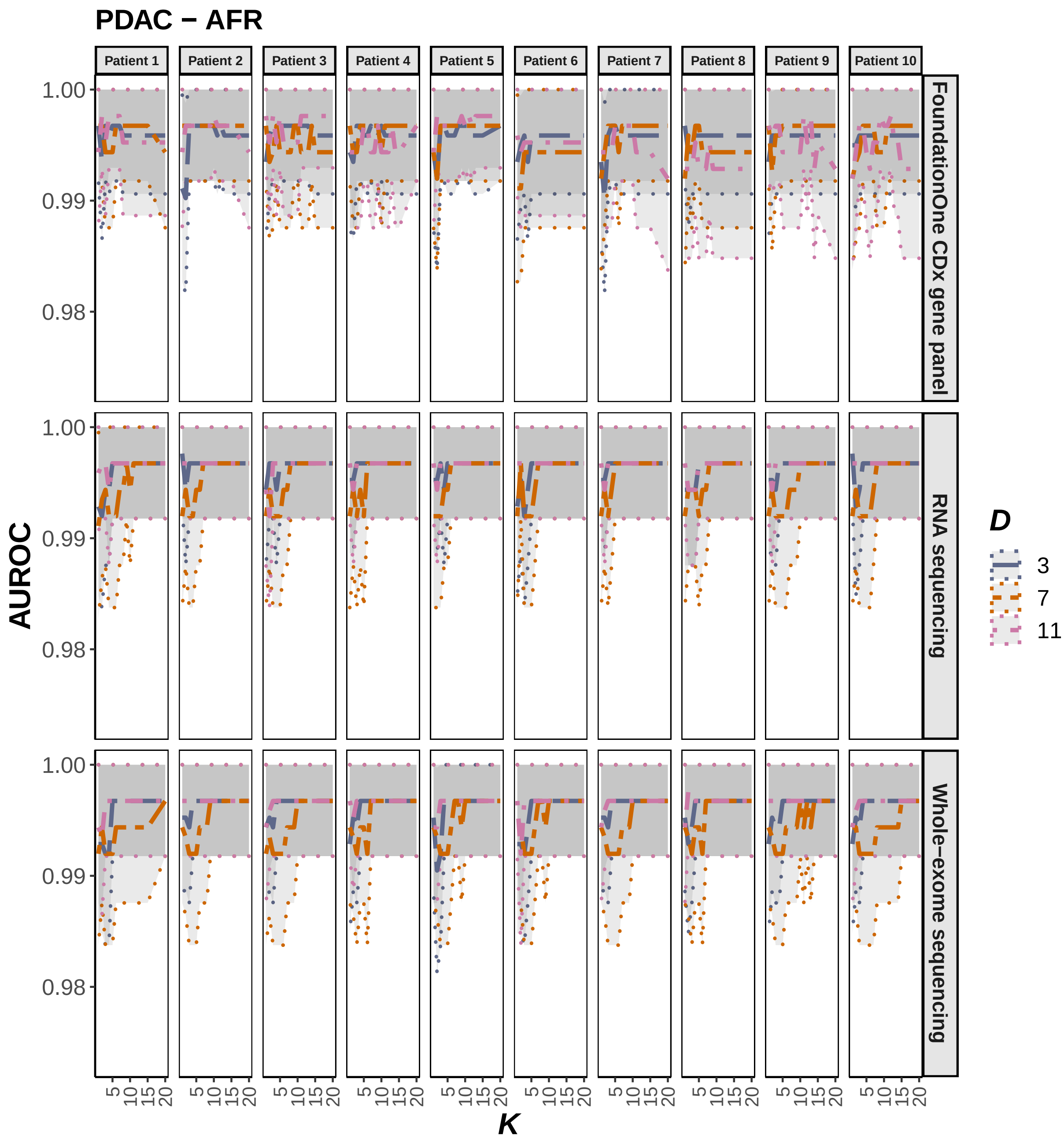

Figure S2. K) Dependence of AFR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 PDAC patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

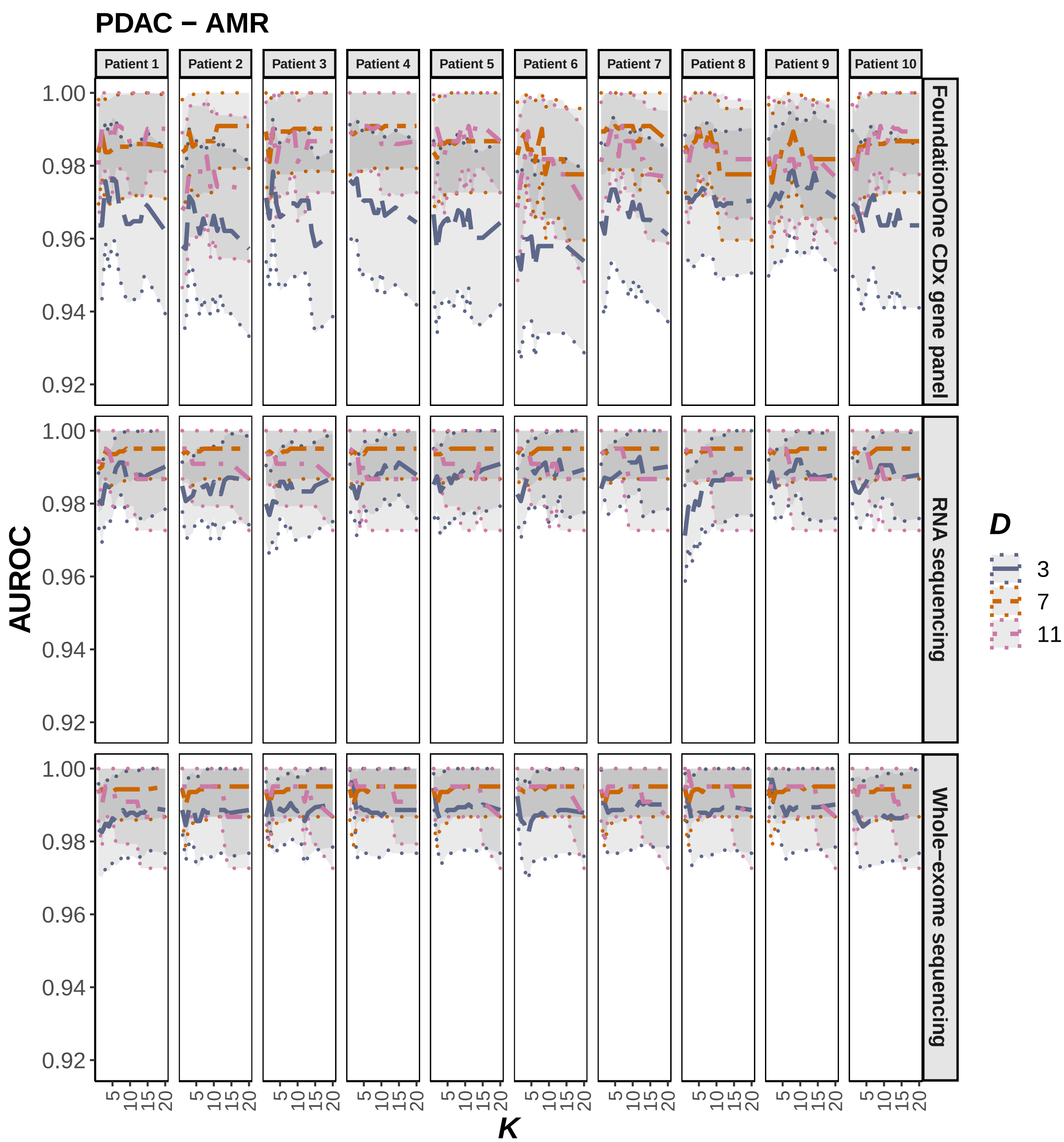

Figure S2. L) Dependence of AMR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 PDAC patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

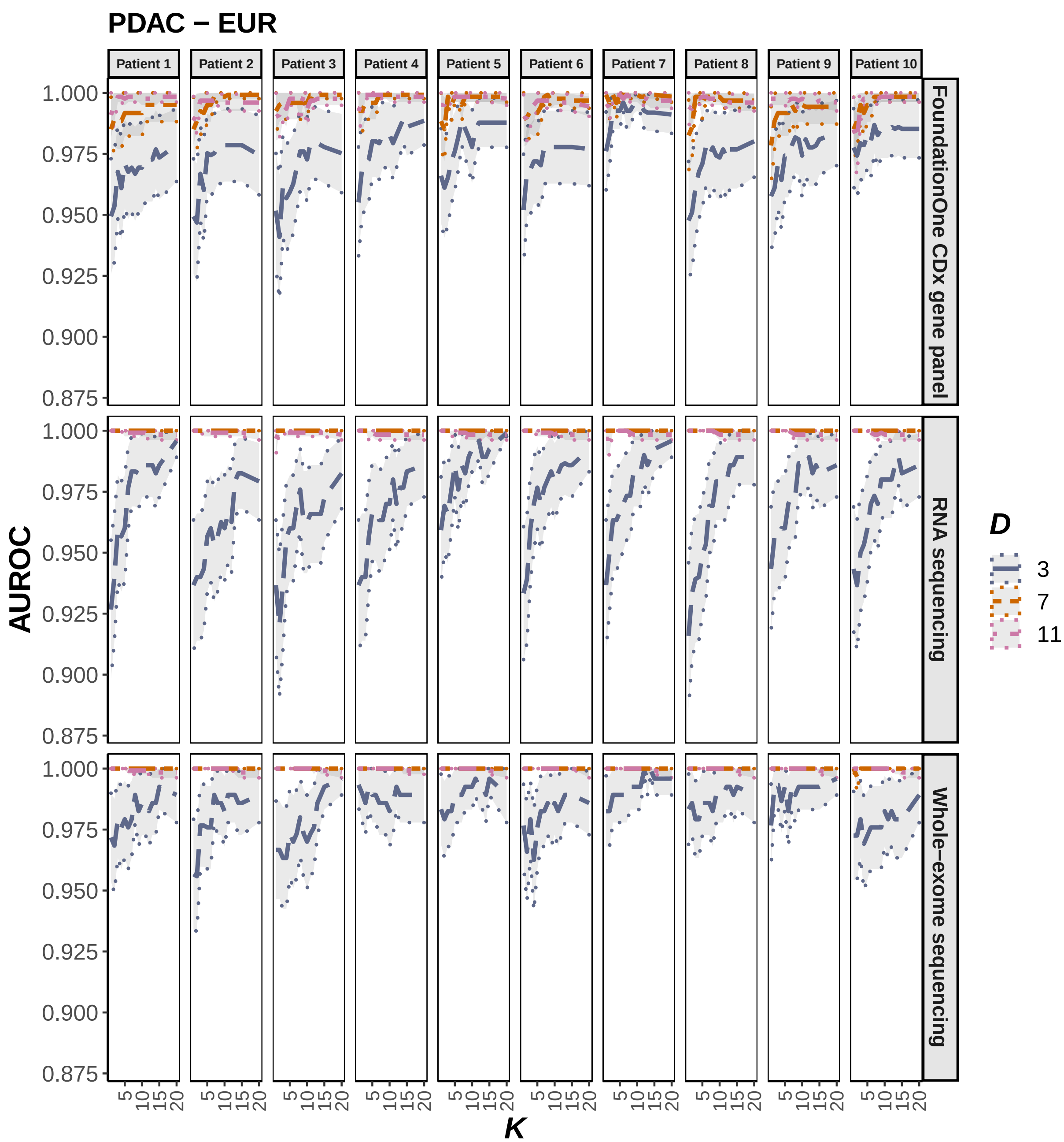

Figure S2. M) Dependence of EUR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 PDAC patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

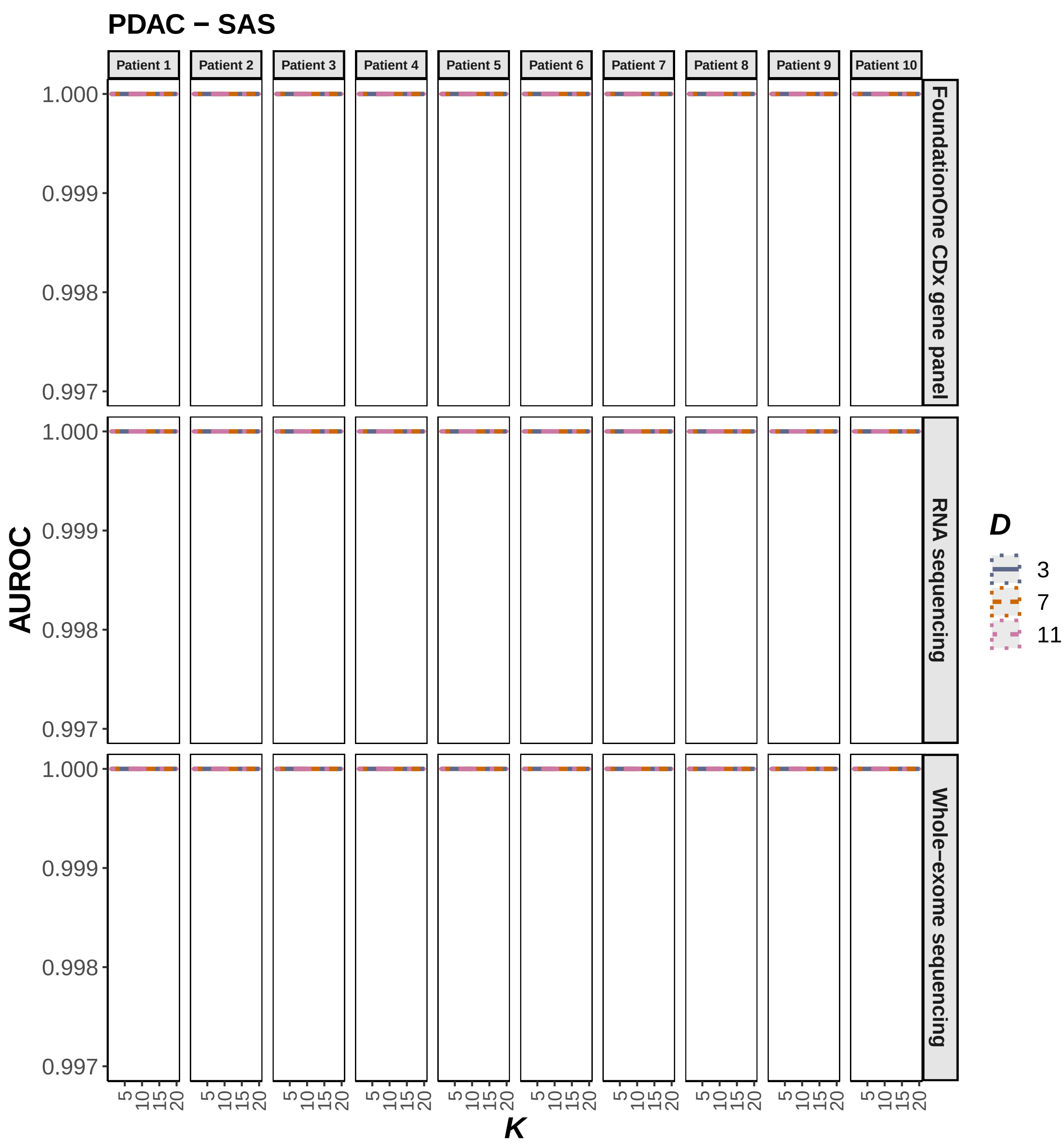

Figure S2. N) Dependence of SAS-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 PDAC patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

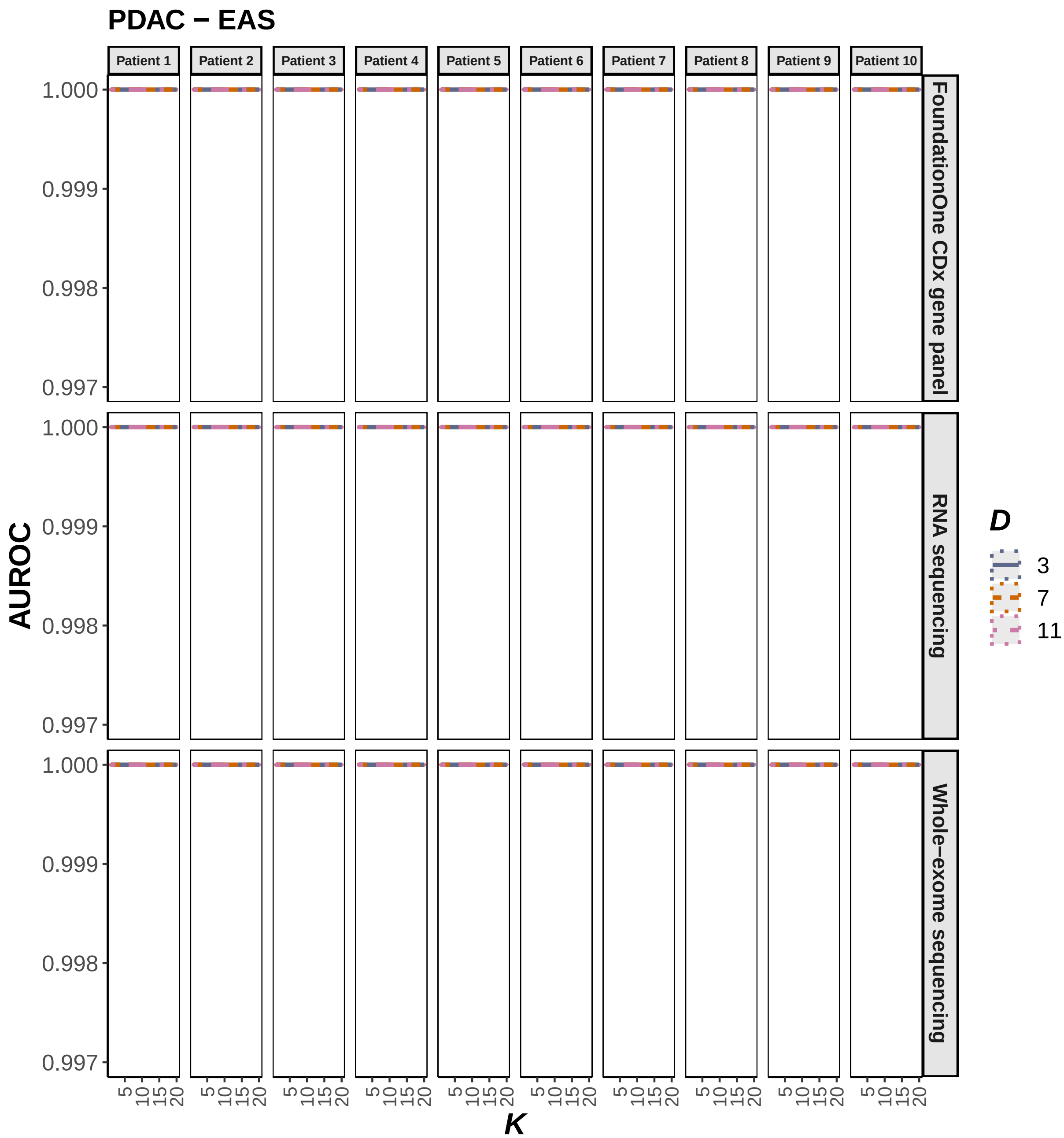

Figure S2. O) Dependence of EAS-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 PDAC patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

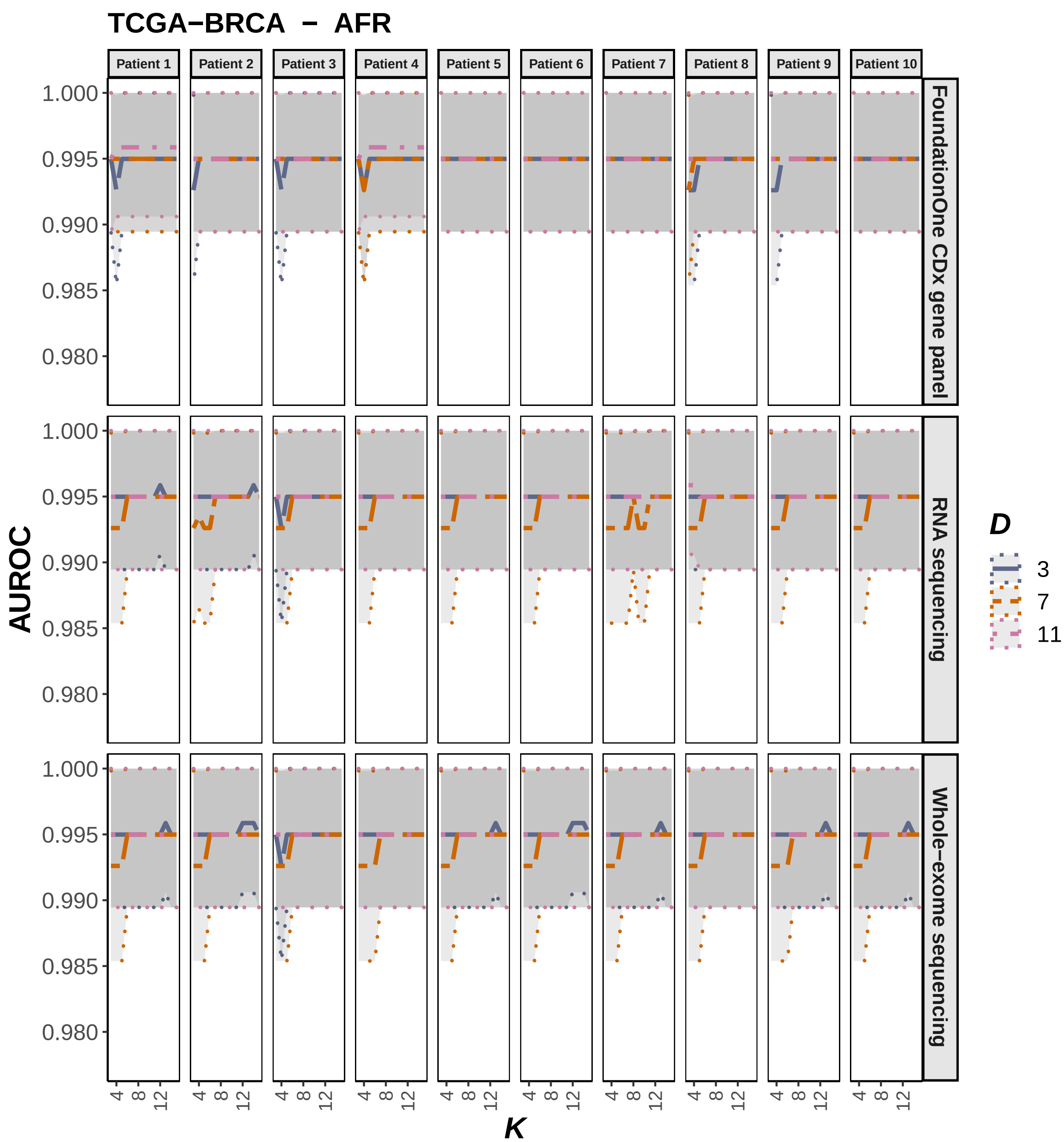

Figure S2. P) Dependence of AFR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-BRCA patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

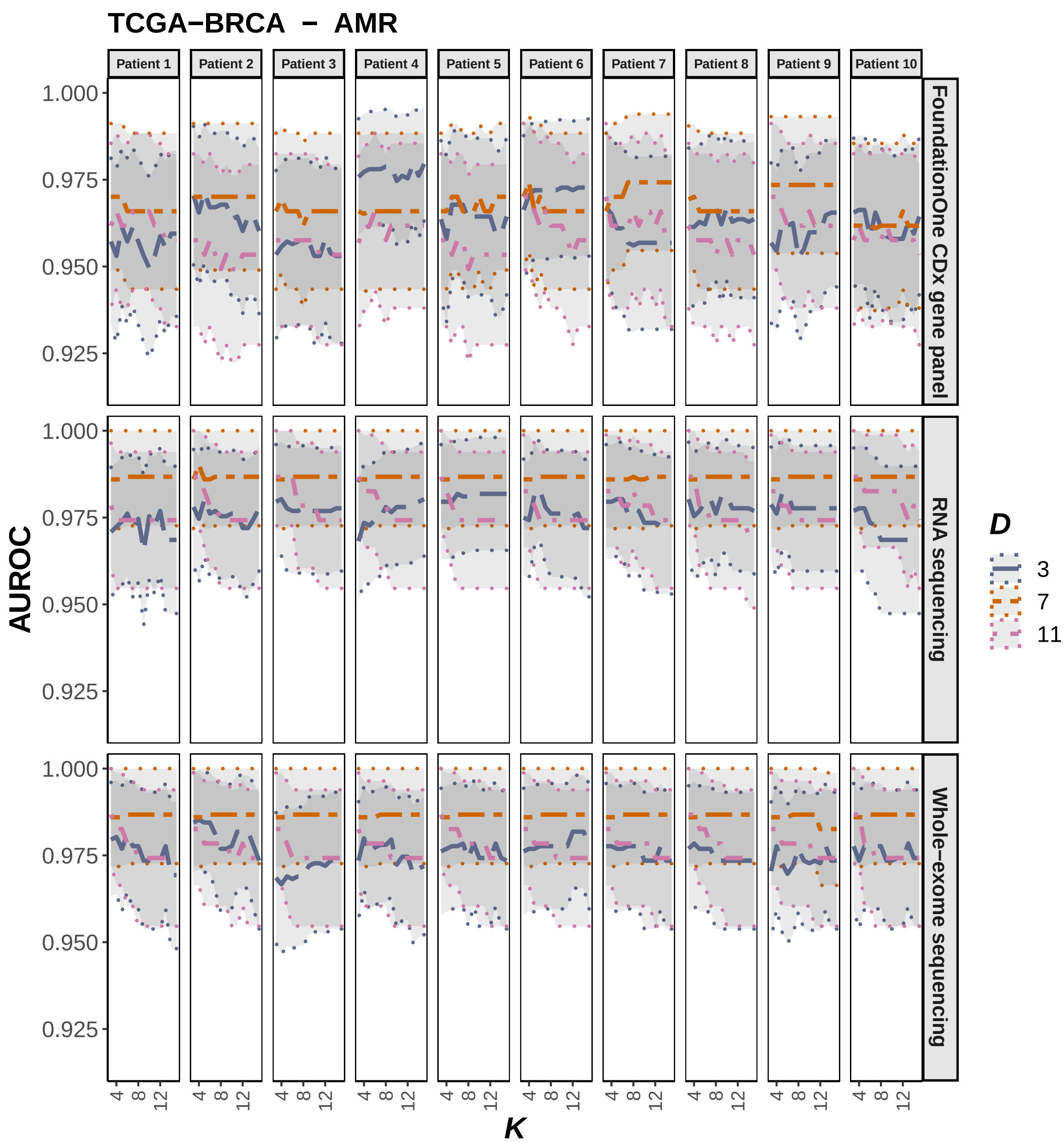

Figure S2. Q) Dependence of AMR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-BRCA patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

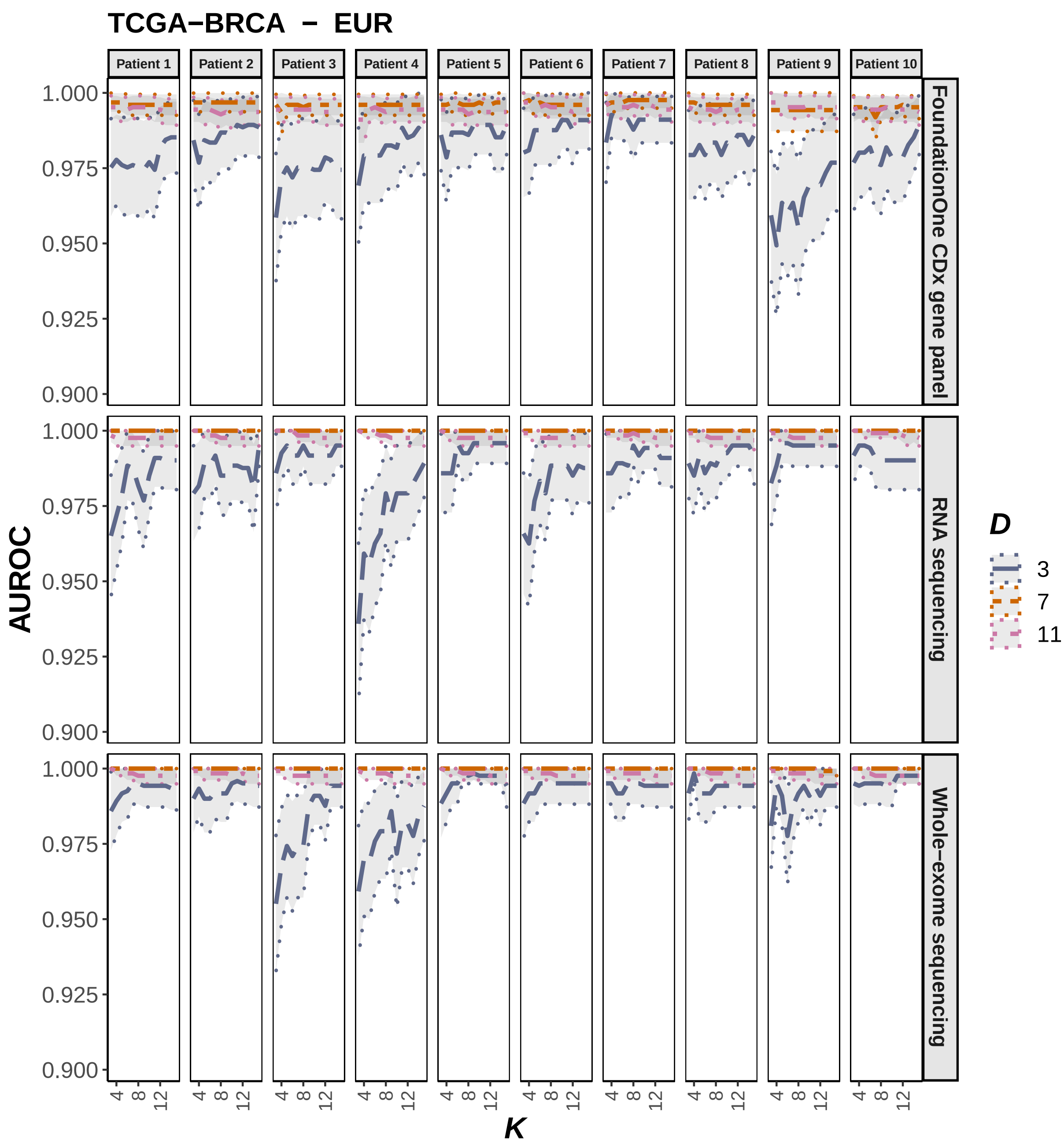

Figure S2. R) Dependence of EUR-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-BRCA patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

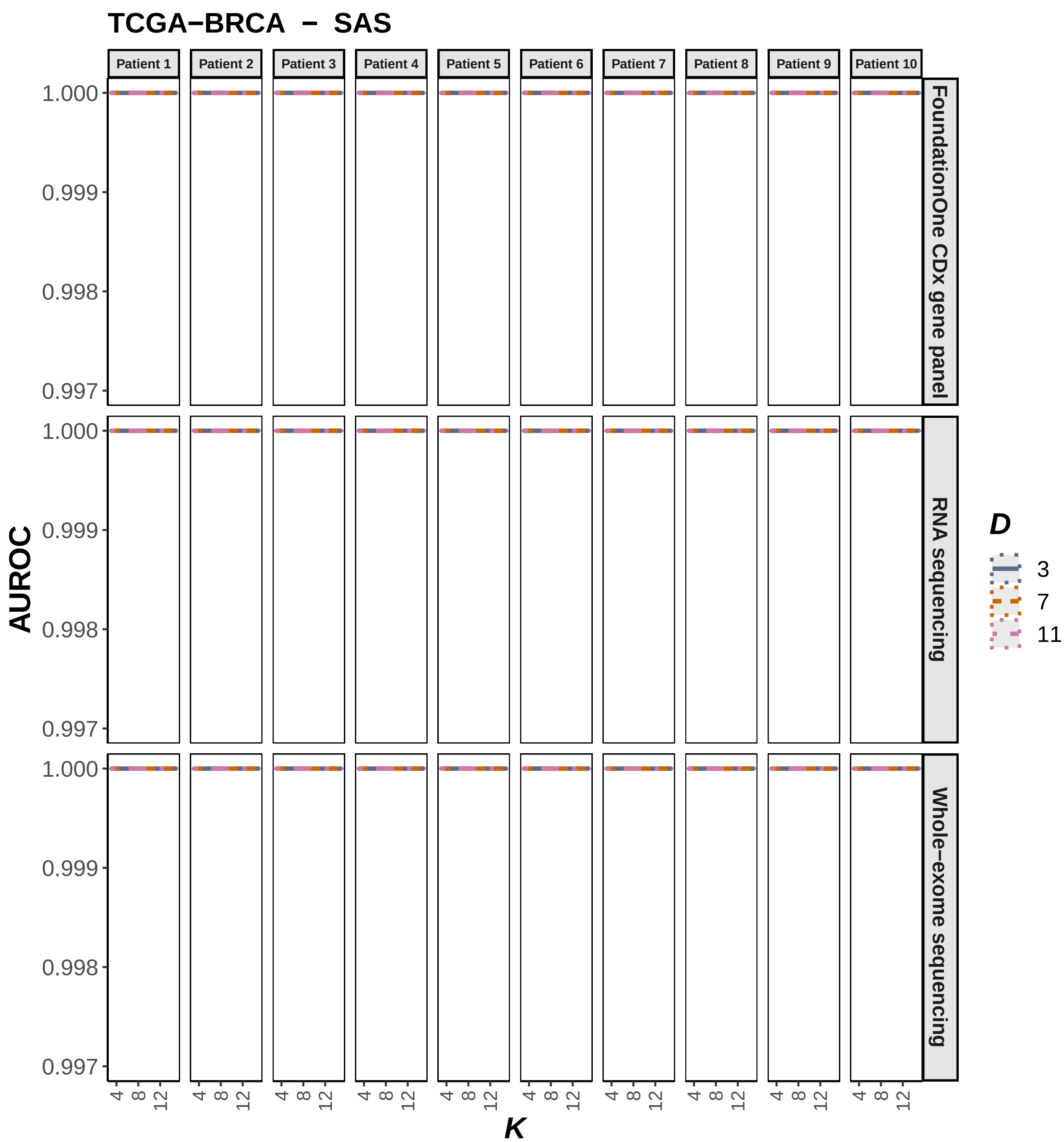

Figure S2. S) Dependence of SAS-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-BRCA patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

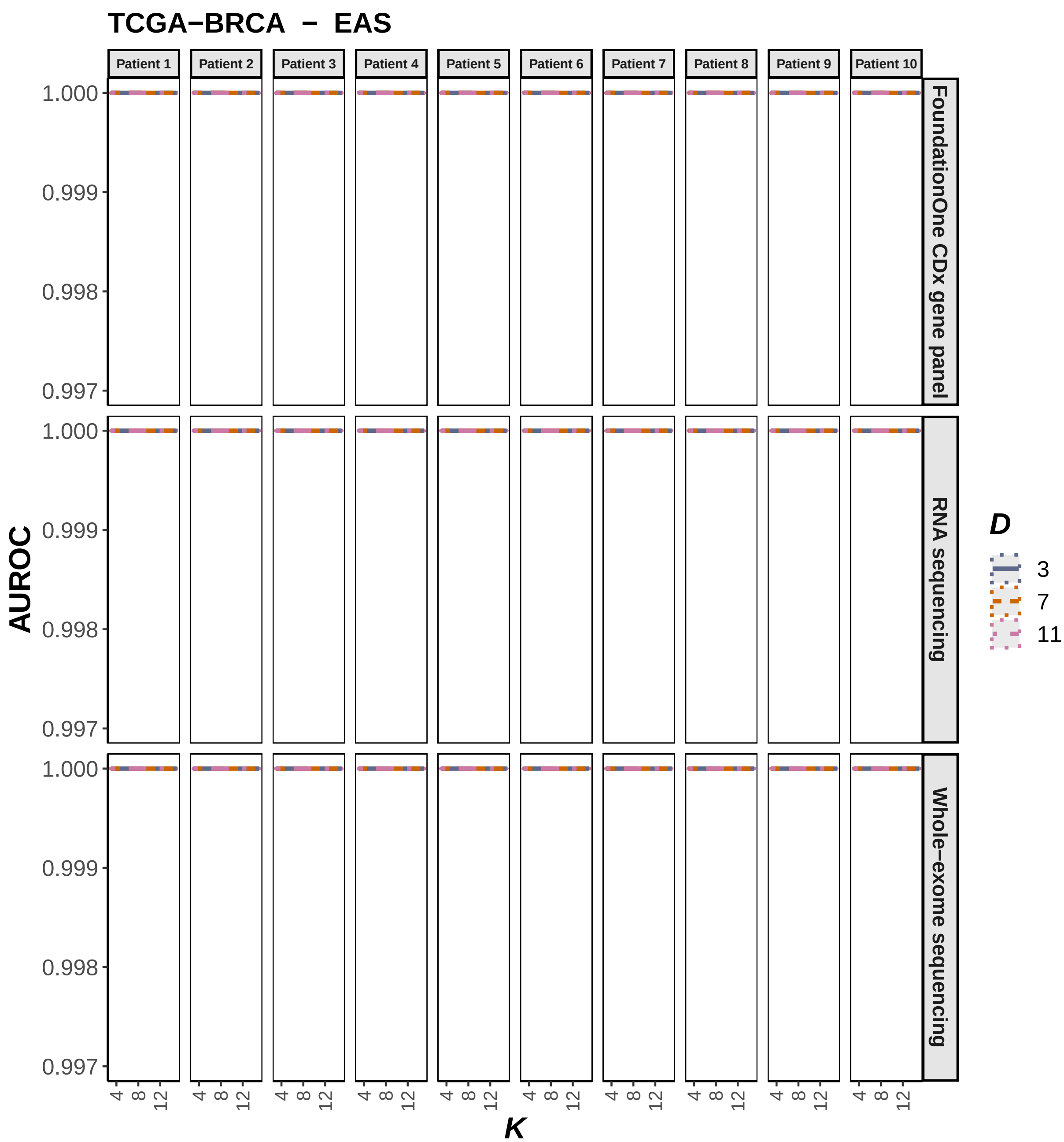

Figure S2. T) Dependence of EAS-specific AUROC on the inference parameters D and K, computed using data synthesis for 10 TCGA-BRCA patients and the three profiling modalities: WES, RNA-seq and FoundationOne® CDx panels. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

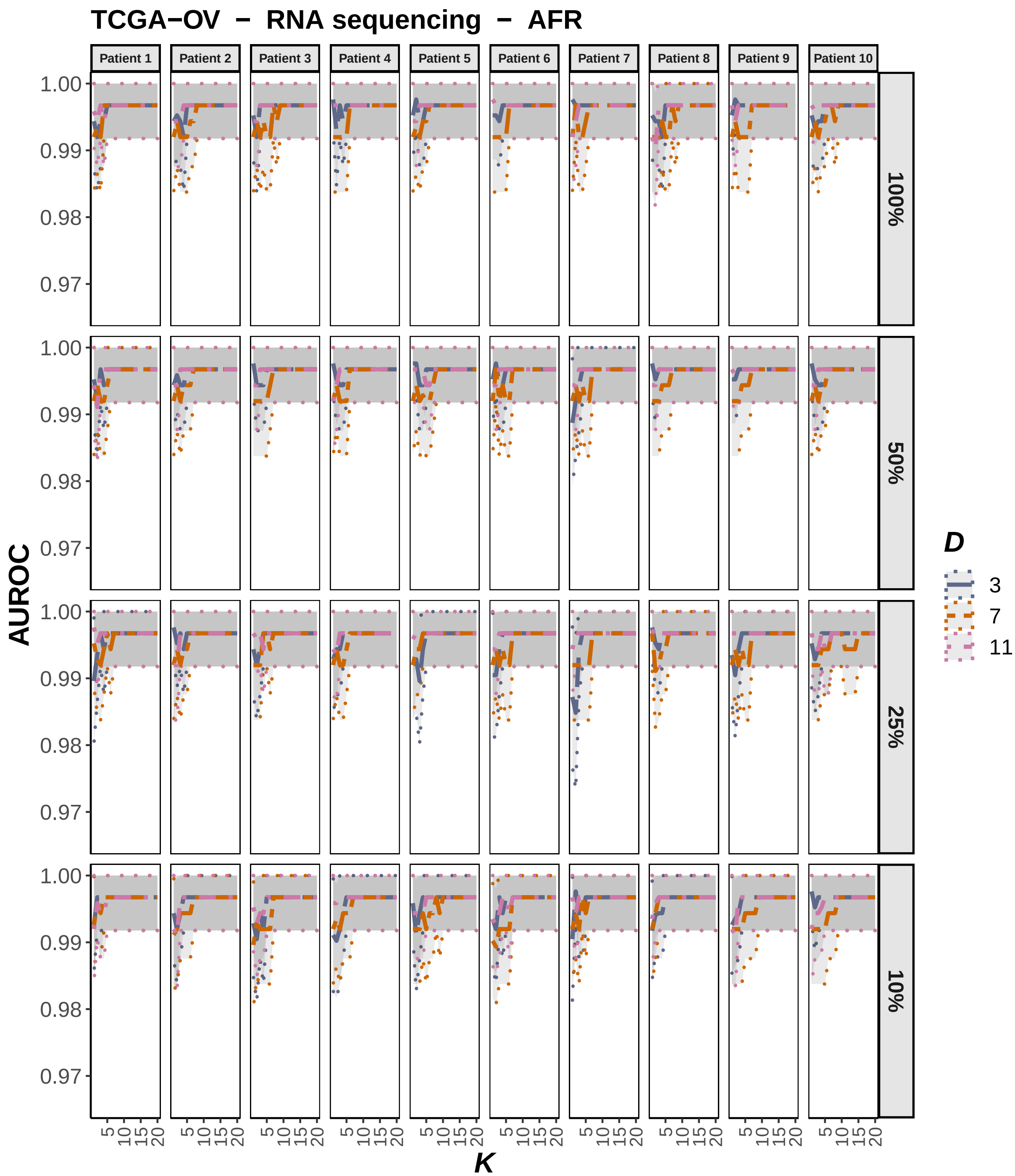

Figure S3. A) Dependence of AFR-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for RNA-seq profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

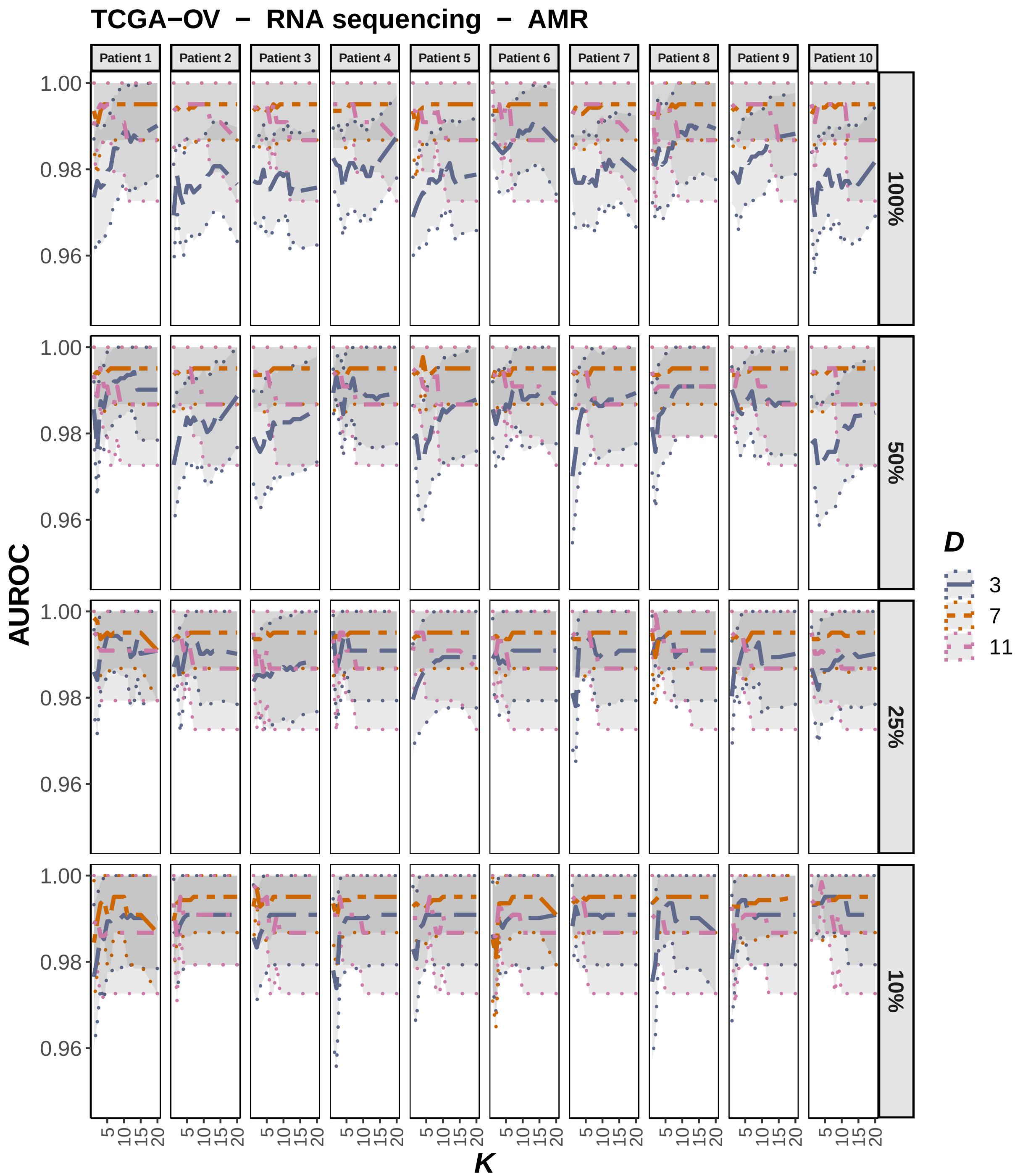

Figure S3. B) Dependence of AMR-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for RNA-seq profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

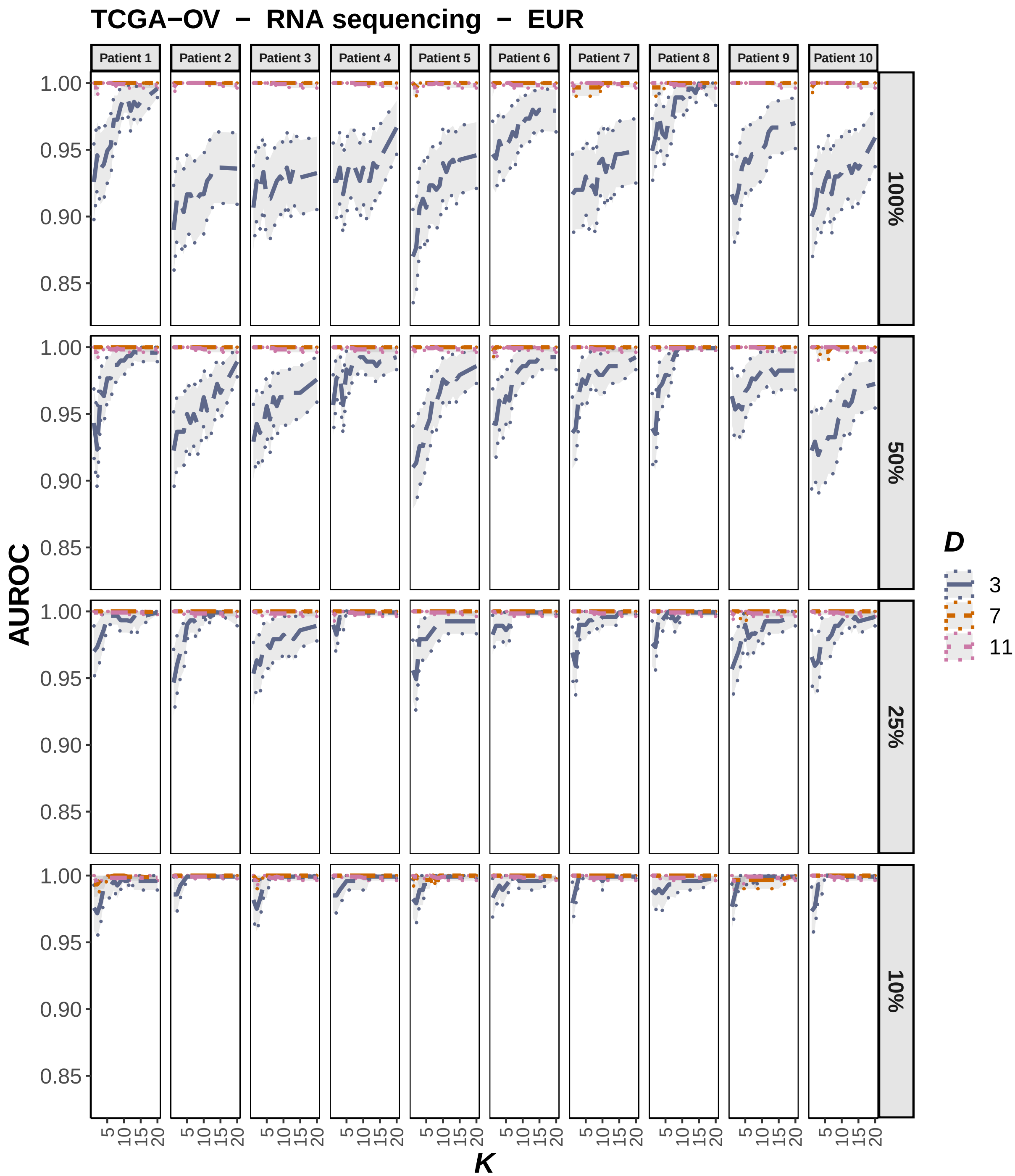

Figure S3. C) Dependence of EUR-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for RNA-seq profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

Figure S3. D) Dependence of SAS-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for RNA-seq profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

Figure S3. E) Dependence of EAS-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for RNA-seq profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

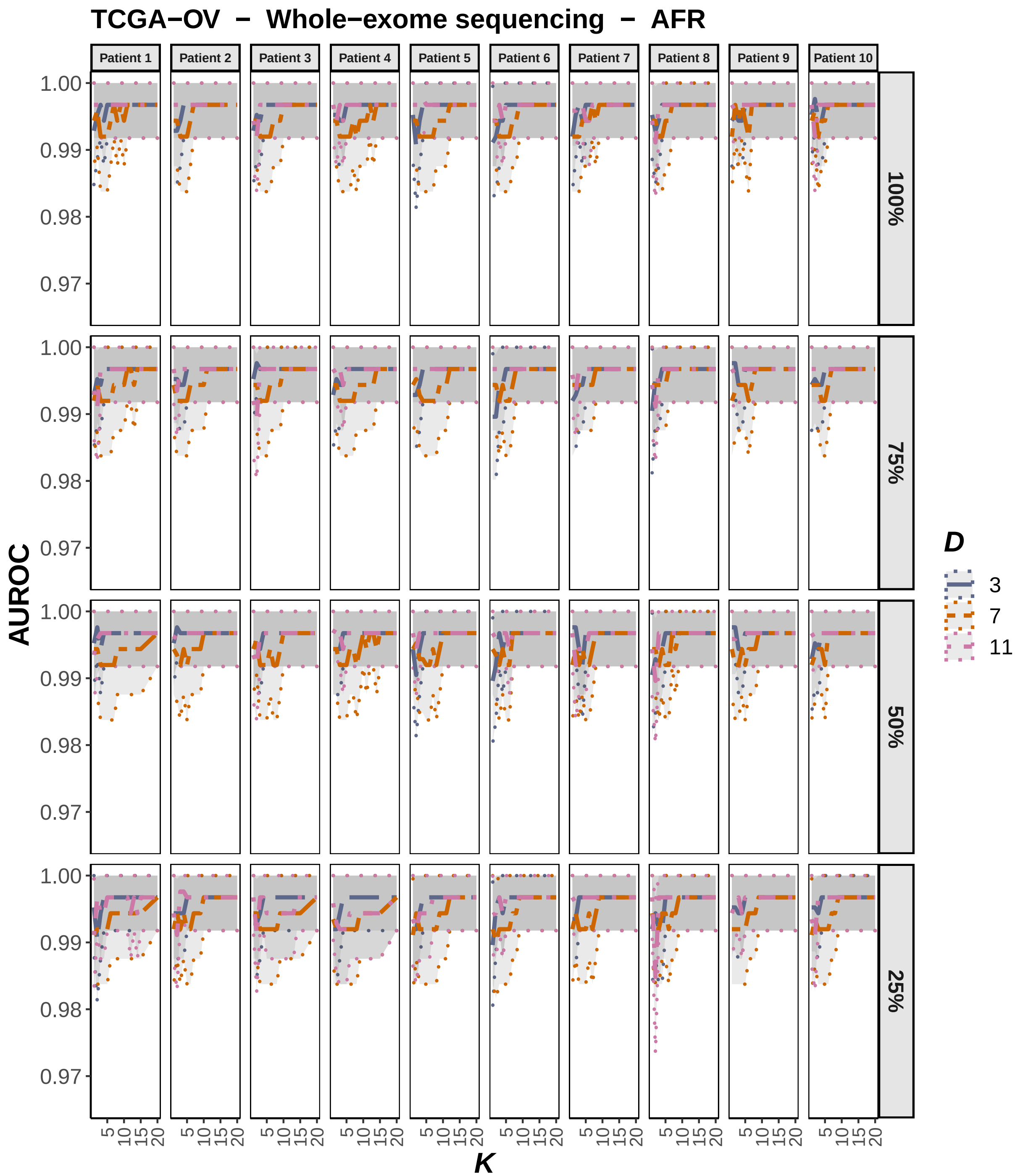

Figure S3. F) Dependence of AFR-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for WES profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

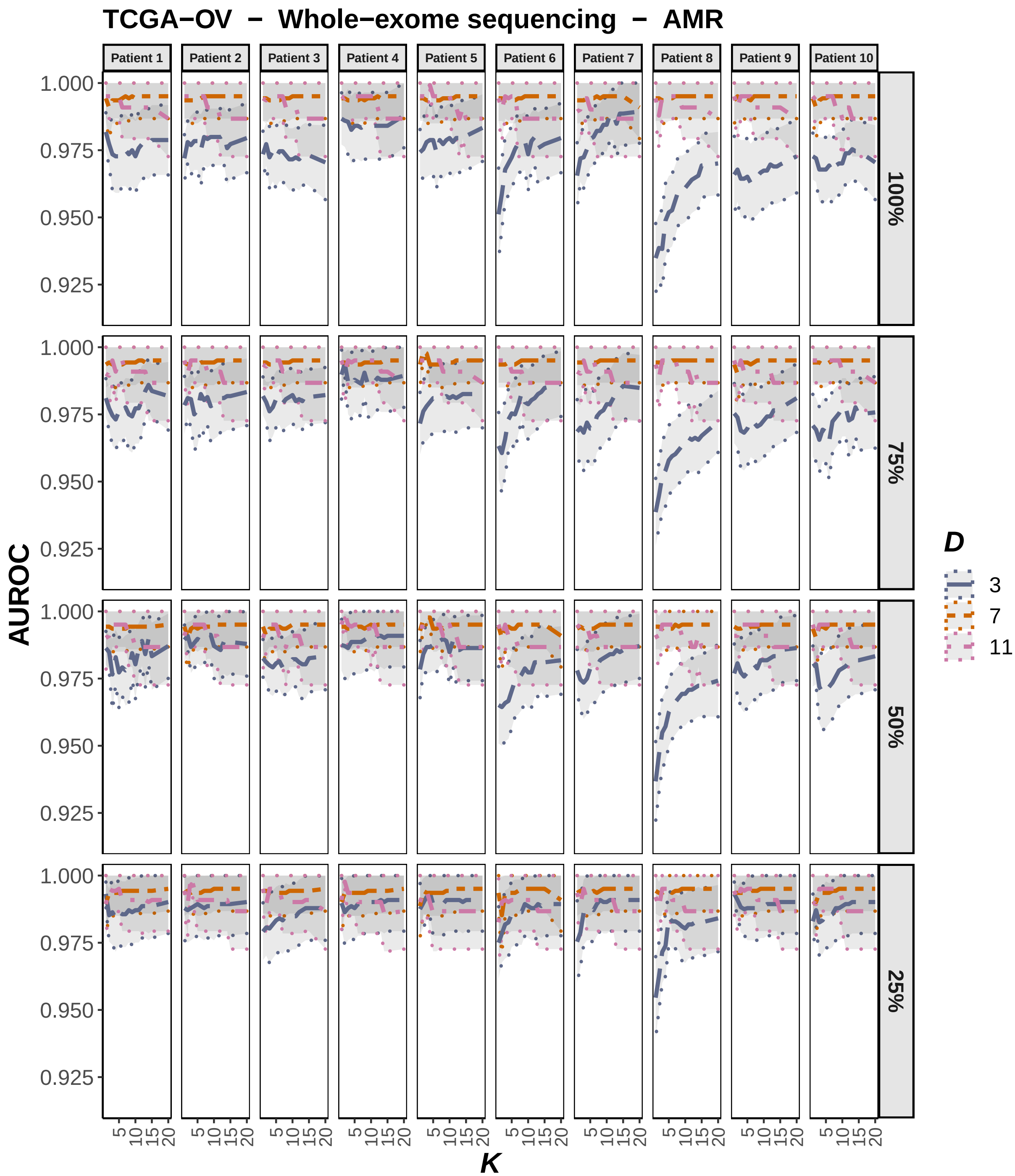

Figure S3. G) Dependence of AMR-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for WES profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

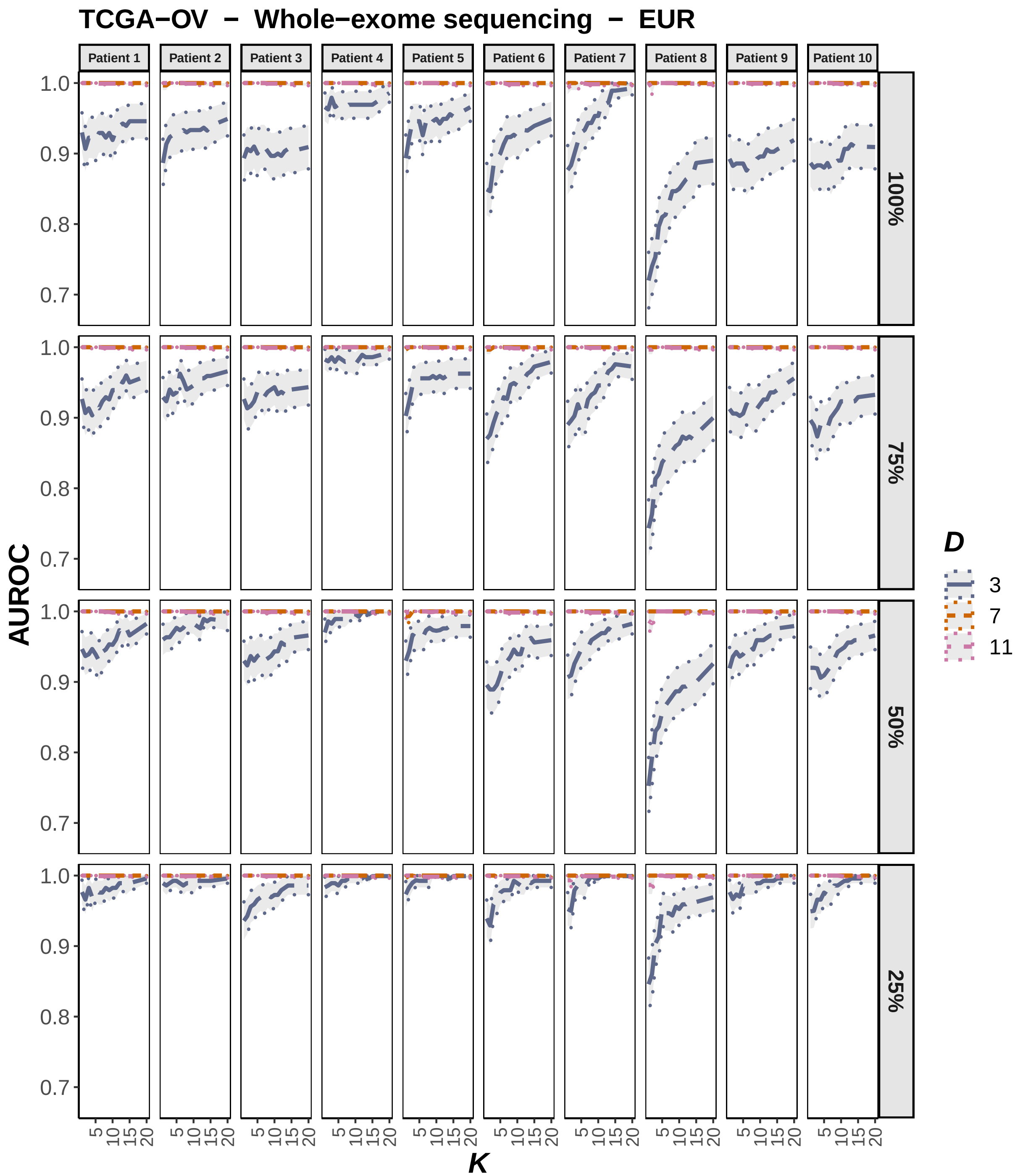

Figure S3. H) Dependence of EUR-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for WES profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

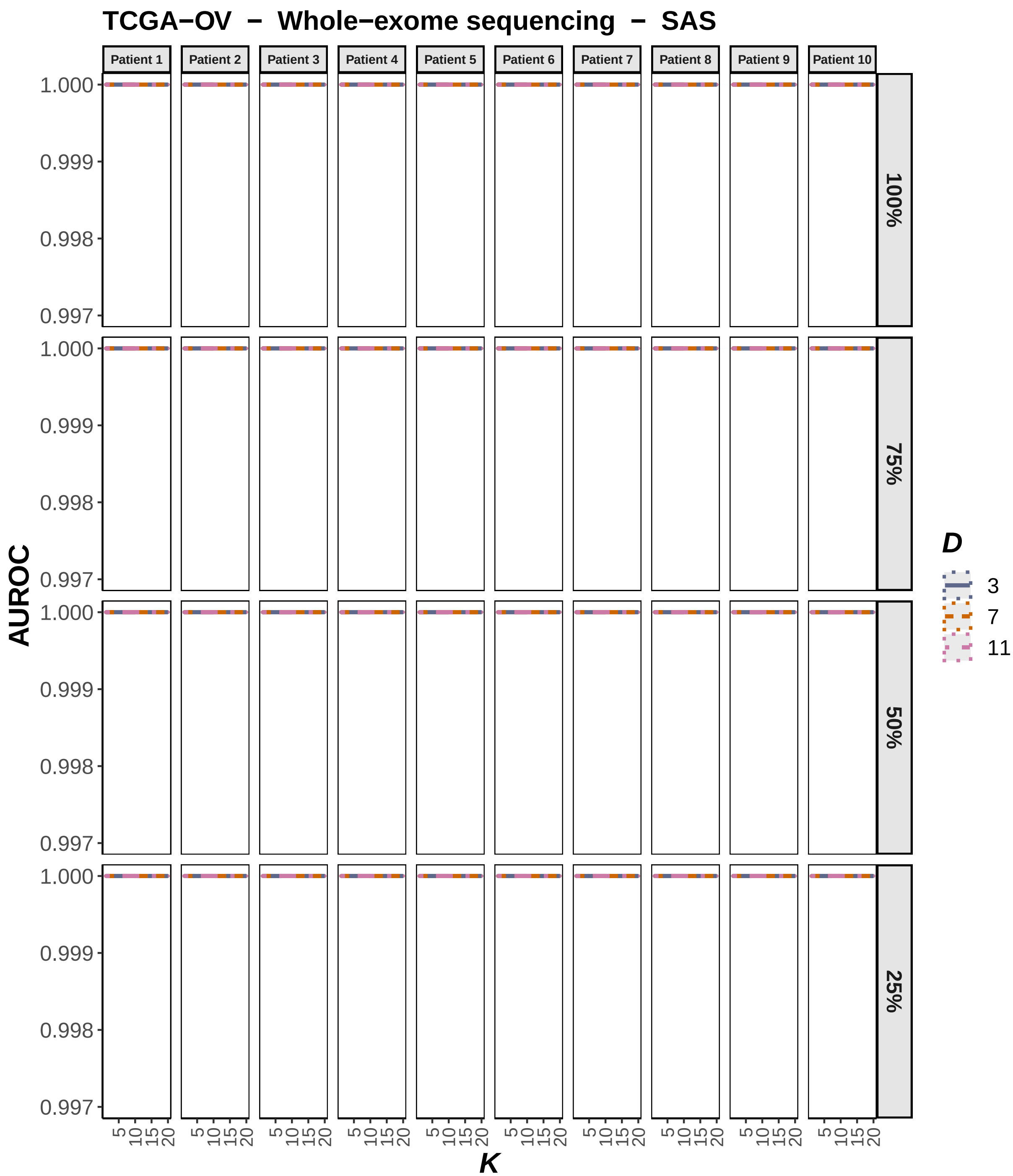

Figure S3. I) Dependence of SAS-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for WES profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

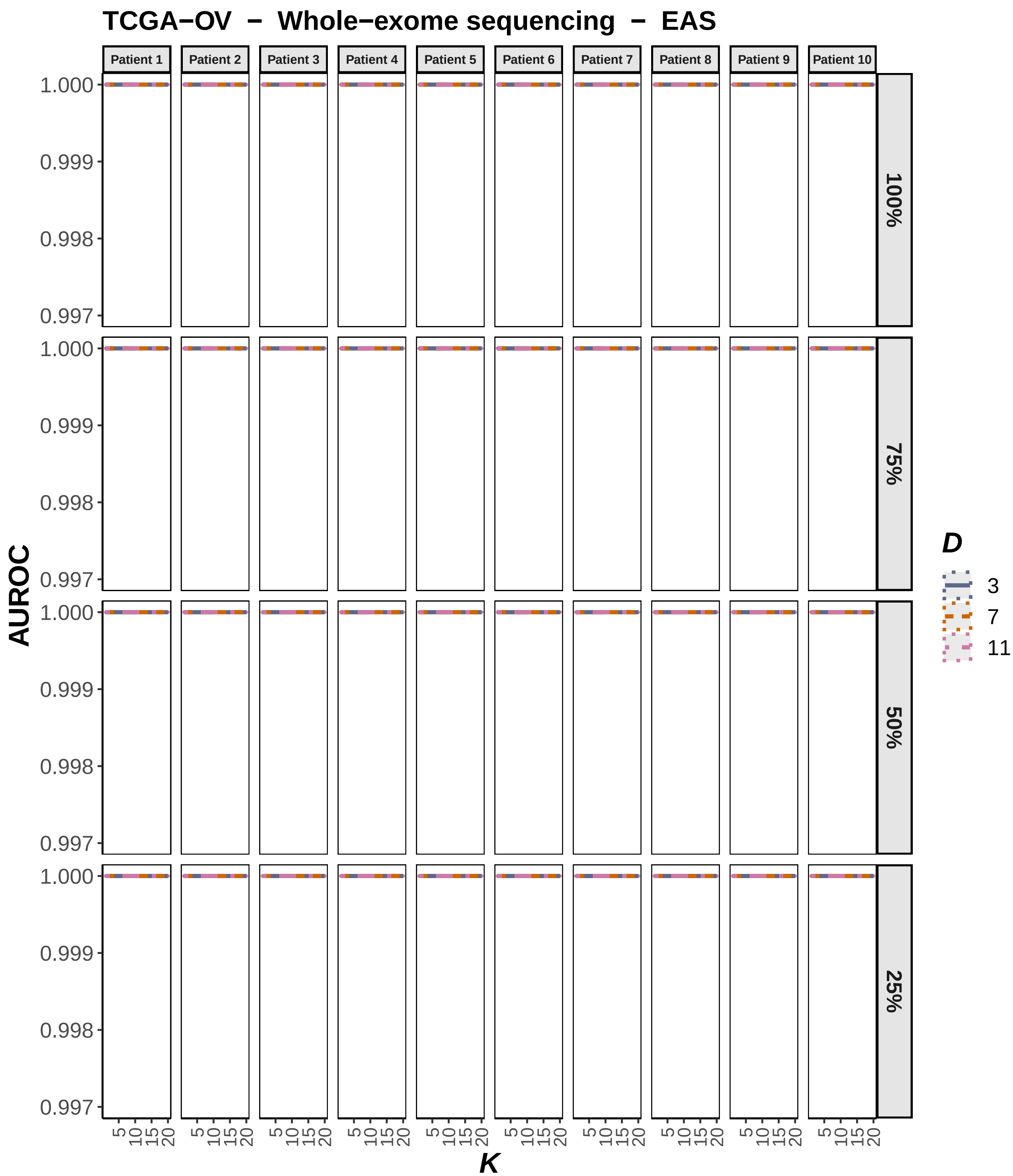

Figure S3. J) Dependence of EAS-specific AUROC on the inference parameters D and K, at the original and reduced sequence coverage values, as indicated by the percentages of the original coverage. AUROC was computed using data synthesis for WES profiles of 10 TCGA-OV patients. The central AUROC values are shown in solid, and the 95% CI in dashed, lines.

References

Carrot-Zhang J, Chambwe N, Damrauer JS, Knijnenburg TA, Robertson AG, Yau C, Zhou W, Berger AC, Huang KL, Newberg JY, Mashl RJ, Romanel A, Sayaman RW, Demichelis F, Felau I, Frampton GM, Han S, Hoadley KA, Kemal A, Laird PW, et al. Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. Cancer Cell. 2020; 37(5):639–654 e6. https://www.ncbi.nlm.nih.gov/pubmed/32396860, doi: 10.1016/j.ccell.2020.04.012.