Detection Rate of Watermarking after Attack as function of Tokens Observed (@ 0.1% FPR)