Supplementary Figure Legends

Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications

Popova T, Manié E, Boeva V, Battistella A, Goundiam O, Smith NK, Mueller CR, Raynal V, Mariani O, Sastre-Garau X, Stern MH

Figure S1. *CDK12* mutations detected using IGV (Integrative Genomic Viewer, (11)) in 2 TCGA samples with highly altered by the middle-scale gains tumor genomic profiles. A. TCGA-24-1551: a splice mutation seen from the WES and fragments of alternatively spliced transcripts from the RNA-seq; B. TCGA-61-2000: a deletion of 19 bp seen from the WES and a drop of coverage around this deletion in the RNA-seq.

Figure S2. Evaluating *CDK12* status in the TCGA cohort.
A. Two cases (TCGA-09-1667 and TCGA-30-1866) with *CDK12* promoter methylation. Red stars point to the tumors with extreme low value of *CDK12* expression and high level of promoter methylation. Other extreme low expression in CDK12 (blue stars) point to truncating and frameshifting mutations in *CDK12*. B-C. Two tumor genomic profiles with missense mutations in *CDK12* kinase domain, which do not display massive middle-scale gain pattern of alteration, characteristic of *CDK12* mutated ovarian tumors. B. TCGA-25-2392 (p.R882L) is mutated in *BRCA1* and the genome displays highly altered profile with aneuploidy and intensive inter-chromosomal breaks. C. TCGA-59-2351 (p.K975E) is mutated in *BRCA2* and the genome displays aneuploidy and intensive inter-chromosomal breaks.

Figure S3. Mutations in *CDK12* found in 4 tumors with the intensive gains in the tumor genomic profiles from the in-house cohort of serous ovarian carcinomas. Three cases were further studied by WGS: MP2, MP3, and MP4.

Figure S4. Examples of chromosome 2 from MP2 (A), MP3 (B) and MP4 (C) tumors from the in-house cohort, where SNP-array and WGS detected alterations are shown. Green horizontal bars at the WGS profile correspond to tandem duplications found by WGS with high confidence (number of read pairs supporting alterations are >20 and had top mapping quality); red triangles show positions, where SNP-arrays report a breakpoint not found by WGS (no structural alteration detected within 0.5 Mb from the breakpoint); black stripes point to translocations. This figure illustrates the rather exhaustive list of alterations found by both techniques. SNP-arrays and WGS displayed a high consistency: more than 70% of alterations detected by SNP-arrays were supported by WGS structural rearrangements and more than 60% of structural alterations detected by mate-pair sequencing were supported by breakpoints detected in the SNP-array profiles. AD: allelic difference profile from SNP arrays; WGS: structural alterations found by NGS; CN: copy number profile from SNP arrays and recognized absolute copy number with 2 and 4 copy levels indicated.

Figure S5. The size of microhomology at the breakpoints of tandem duplications (A) and the size of TDs (B) in 3 sequenced in-house ovarian tumors with *CDK12* mutations.

Figure S6. Copy number alteration profiles of eaPEO14 primary tumor (A) and PEO14 cell line (B) described in (20) as having a tandem duplicator phenotype. The regular small scale gains in the tumor genomic profile were annotated by Ng and colleagues (20) as tandem duplications based on the WES and extensive validation of breakpoints by Sanger sequencing

Figure S7. *CDK12* mutation in PD3722a sample. A. Copy number profile obtained from WGS data by Control-FREEC software (13) indeed showed increased numbers of interstitial gains in PD3722a B. Manual inspection of *CDK12* by IGV genome browser (11) showed the deleterious nonsense mutation in *CDK12* (chr17:37619135, c.811A>T/p.R271X). C. The amplification of the *CDK12* locus in PD3722a allowed us to discover a nonsense mutation of this gene despite the overall low sequencing coverage. D. Size of validated TDs for PD3722a sample (EGAS00001000155) (25) followed the distribution obtained for the CDK12 inactivated cases; although TD detection from the low-coverage WGS technique was probably incomplete.

Figure S8. Tandem duplications in 25 TCGA cases with WGS. A. The number of TDs detected in 25 tumors shown in decreasing order. Tandem duplications in these series of tumors were found in each case with a median 23 TDs per sample (range: 2-346, median absolute deviation: 7. B. TD size distribution in *CDK12* mutated case (TCGA-24-1466) and in the next abundant in TDs tumor (TCGA-13-1487). C. SNP-array genomic profile of TCGA-13-1487 case showing frequent small-scale gains. This analysis confirmed the structural specificity of *CDK12*-mutated tumor and revealed one case with high frequency of TDs mainly of less than 1Mb in size.

Figure S9. Detecting tumors with tandem duplication (TD) phenotype in the TCGA breast and ovarian cancer cohorts. A. The number of interstitial gains of less than 1.5Mb in size was calculated from SNP array profiles. Tumors with presumable TD phenotype are marked by the yellow ellipse. These tumors were explored manually and 13 cases were selected. Cases with CDK12 TD-plus phenotype were marked by red (ovary) and green (breast) circles. B. 10 ovarian and 3 breast tumors with available sequencing data were selected for a mutational analysis. Two cases marked in blue were attributed to CDK12 TD-plus phenotype, but no *CDK12* mutation was found in these tumors.

Figure S10. TD-plus phenotype in breast and prostate cancers. A. Copy number alteration profile of TCGA-A2-A04U primary tumor from the TCGA Breast Invasive carcinoma cohort: the only case out of 760 breast tumors showing similar level of interstitial gains as tumors with CDK12 TD-plus phenotype in ovarian cancer. However, no *CDK12* mutation was found. The case was marked in green on the Figure S8. B. Analysis of the TCGA cohort of prostate adenocarcinoma (PRAD). 7 cases with TD-plus phenotype were detected among 407 cases analyzed. *Mutations reported in cBioPortal.

Figure S11. GC content and mutation rate in TDs. A. GC content in the genomic segments affected by tandem duplications summarized by the TD peak density (the number of tumors having TD in the genomic segment, such as shown in Figure S12C). B. Mutation rates per bp in TD and not TD segments obtained for the TCGA-24-1466 case. The rates are not significantly different and not increased in the genome segments that underwent tandem duplication in *CDK12* mutated tumors.

Figure S12. Recurrence analysis of TDs in the tumors with CDK12 TD-plus phenotype. A. The rainfall plot for TDs in 4 *CDK12*-mutated ovarian tumors with available WGS (MP2, MP3, MP4 and TCGA-24-1466). The distance between two consecutive TDs was calculated as the distance between the centers of TDs. Each point represents a TD and the value in vertical axis corresponds to the distance to the next TD along the genome. B. Random sampling (black curve) gives similar between-center distance distribution as the actual one (red histogram) observed in the tumors and illustrated in panel A. C. TD density per tumor in the SNP-array profiles (17 cases, shown at the top) and in WGS (4 cases shown at the bottom). In the SNP-array profiles, all interstitial gains less than 7 Mb were considered, the copy number gains were adjusted for tumor ploidy. The genomic location of the most prominent peaks (6-7 TDs) is indicated. Among WGS peaks at 1p34, 1q41-42, 2p23, 3q26 and 6p21, only 3q26 was supported by the SNP array summary profiles.

Figure S13. Replication time measured for HeLa cell line and the distance to CTCF binding sites with respect to tandem duplications found by WGS. A. Tandem duplications were subdivided in two subgroups depending on the replication time in the central part of a TD (early or late). TDs with early replicating center are shown in the left panel and TDs with late replicating center are shown in the right panel. Replication time is indicated by the color scale. No particular pattern of replication timing on the TDs is observed. TDs are equally represented by early and late replication timing pattern. B. The histogram of the distance from CTCF to TD breakpoint is shown (red bars: observed in the cell lines; white bars: random sampling).

Figure S14. Tandem duplications in *CDK12* mutated tumors and gene loci. A. Random sampling of TDs of the same sizes as TDs in the list obtained from WGS of 3 in-house cases mutated for *CDK12*: MP2, MP3, MP4. Top panels show the frequency of appearance of the random TD breakpoints within gene (left panel) or expressed gene (right panel) loci. Blue stars indicate the actual proportion of TDs breakpoints obtained for the sequenced tumors. Low panels show proportion of TD center to intersect with gene (left panel) or expressed gene (right panel) loci. Mean proportions and standard deviations for the random sampling are indicated. B. All TDs validated by split-reads were sorted according to the size and the

top 60% expressed genes are marked by color-code reflecting quantiles of the expression level. Each horizontal line represents one centered TD. The appearance of TDs in *CDK12* inactivated tumors is unlikely to be primary associated with gene or expressed gene loci.

Figure S15. Three genomic signatures of HRD on the set of tumors with CDK12 TD-plus phenotype (A) and on all TCGA dataset of ovarian tumors (B). While showing consistent behavior on the full set of ovarian tumors, prediction of HRD in tumors with CDK12 TD-plus phenotype are particularly not consistent among HRD signatures. The CDK12 TD-plus phenotype precludes robust estimation of genomic HRD. Extensive filtering of interstitial gains is instrumental in obtaining consistent HRD evaluation.

Figure S16. Copy number genome profiles in tumors with CDK12 TD-plus phenotype. LSTs numbers were found close to the thresholds (defining HRD) in the first 5 cases. The blue line shows absolute copy number profile; the green under-lines indicate the large segments, defining LST at the copy number break. Intensive copy number variation besides the small scale gains is characteristic of homologous recombination deficiency.

Figure S17. *CDK12* mutated tumors and mesenchymal subtype in the TCGA (A) and In-house (B) cohorts. The gene set representing stromal and immune components of the normal cells admixture to tumor sample was taken from (26), because the mesenchymal subtype was mainly defined by the stromal cell admixture. This gene set was used for consistency mainly because of the in-house cohort, which is small and heterogeneous for stable and reliable classification into four molecular subtypes. Principal component analysis was applied and the data were visualized in the two first principal components (PC1, PC2), where each point represents a tumor; tumors with CDK12 TD-plus phenotype are designated by black stars. A. Each tumor is designated by the subtypes and the corresponding color: Me: Mesenchymal (blue), Im: Immunoreactive (green), Di: Differentiated (red), Pr: Proliferative (violet); B. Each number represents a tumor by its level of expression of collagen, one of the main determinants of the Mesenchymal subtype.

Figure S18. Specific molecular features of tumors with CDK12 TD plus phenotype and differential analysis. A. Tumors with CDK12 TD plus phenotype and expression of

CDKN2A/PTEN (left panel) and *CCNE1/CCND1* (right panel). Each tumors is designated by the subtype abbreviation: I: Immunoreactive, D: Differentiated, P: Proliferative; tumors with CDK12 TD-plus phenotype are designated by red stars. Tumors with CDK12 TD-plus phenotype are never depleted for *PTEN*, however, this is not significant due to rare incidence. B. P-value distribution in comparison of the gene expression of 17 tumors with CDK12 TD-plus phenotype versus 57 matched cases on the TCGA Affymetrix platform. Even though comparing the small sample sets of tumors, the p-value distribution shows significance of the differences found and indicates FDR (false discovery rate) around 0.25 (60/222). C. Validation of the gene set (222 genes) obtained from comparison on the TCGA Affymetrix platform (described above in B). The direction of the fold change (FC) was identical on Affymetrix and RNA-seq TCGA comparisons in 97% of genes. With this RNA-seq p-value showed at least minimal significance ($p<0.05$) in 60% of genes. In-house Affymetrix (3 *CDK12*-mutated versus 34 matched cases comparison) validated 56% of genes on the level of fold change and 42% of them showed <0.05 significance. FC: genes with consistent direction in fold change are only counted.

Figure S19. Set of genes correlated to *CDK12* in the TCGA and in-house cohorts. A. First principal component (PC1) calculated for the set of genes, which are correlated to *CDK12* ($|r| > 0.5$) versus expression of *CDK12*. *CDK12*-inactivated samples are designated by orange (missense mutation), red (truncating mutation) or green (promoter methylation) stars. Drop out of truncating mutations probably reflects nonsense mediated decay (NMD). Diverse PC1 coordinates of *CDK12*-inactivated cases (PC1 is not down-regulated following *CDK12* expression) probably evidence adaptation to the inactivated gene. B. The set of genes which have positive correlation (>0.6) with PC1 (blue bars) are enriched with the large genes (total gene size was considered) compared to all set of expressed genes (white bars).

Figure S20. Expression of *ATM* and *CSTF3* and *RPRD1A* in a *CDK12* context. A. First principal component (PC1) calculated for the set of genes, which are correlated to *CDK12* ($|r| > 0.5$) versus expression of *ATM* in RNA-seq (left panel) and Affymetrix (right panel) platforms of the TCGA ovarian cancer cohorts. Linear model ATM~PC1 was built and the residues (CDK12 TD-plus phenotype tumors versus other cases) were compared by the t-test. For the RNA-seq data $p<0.02$ was obtained. B. Expression of *CSTF3* versus relative copy number in 11p13

genomic segment on RNA-seq platform (left panel) and versus *RPRD1A* on Affymetrix platform (right panel). Two genes have direct functional connection with *CDK12*.
CDK12-inactivated samples are designated by orange (missense mutation), red (truncating mutation) or green (promoter methylation) stars.

Figure S21. Alteration profiles of SKOV3 cell line, which showed increased number of interstitial gains (>50) largely (>80%) annotated as tandem duplications by WGS. A. SNP-array copy number profile of SKOV3 cell line from the CCLE (27). B. Zoom-in to chromosome 6 showing tandem duplications detected by WGS. Green horizontal bars at the WGS profile correspond to tandem duplications found by WGS with high confidence (number of read pairs supporting alterations are >20 and had top mapping quality); red triangles show positions, where SNP-arrays report a breakpoint not found by WGS (no structural alteration detected within 0.5 Mb from the breakpoint); black stripes point to translocations. AD: allelic difference profile from SNP arrays; WGS: structural alterations found by NGS; CN: copy number profile from SNP arrays and recognized absolute copy number with 2 and 4 copy levels being indicated. C. Circos plot of structural rearrangements found by WGS (28). Red: TDs; Blue: deletions; Violet: translocations. D. The size of tandem duplications (left panel) and the size of microhomology at the junctions (found for 60/77 TDs) (right panel) for SKOV3 cell line. TD: tandem duplications.

Figure S22. The size of microhomology at the breakpoint junctions of tandem duplications shown as a density plot. SKOV3 cell line displayed feeble bimodality with the most frequent microhomology being at 1bp overlap. *CDK12* mutated cases have more pronounced peak in microhomology size at 2 bp compared to SKOV3 cell line and the collection of TDs from *CDK12*-intact tumors (TDs-7). Chi-squared test was applied to compare 5 distributions and in both comparisons the differences were significant. TD: tandem duplications; TDs-7: tandem duplications collected from 7 ovarian tumors without CDK12 TD-plus phenotype and some TDs detected with the sequence at the breakpoint junction (Table S3).