Targeted sequencing of *BRCA1* and *BRCA2* across a large unselected breast cancer cohort suggests one third of mutations are somatic

Christof Winter[†], Martin P. Nilsson[†], Eleonor Olsson, Anthony M. George, Yilun Chen, Anders Kvist, Therese Törngren, Johan Vallon-Christersson, Cecilia Hegardt, Jari Häkkinen, Göran Jönsson, Dorthe Grabau, Martin Malmberg, Ulf Kristoffersson, Martin Rehn, Sofia K. Gruvberger-Saal, Christer Larsson, Åke Borg, Niklas Loman, and Lao H. Saal

Supplementary Methods

Patient cohort and samples

Patients diagnosed with invasive breast cancer who were scheduled for surgery in Malmö, Sweden, during the years 2007 through 2009 were asked to participate in the population-based All Breast Cancer in Malmö (ABiM) study by agreeing to donate a blood sample for research purposes and to consent the use of blood and tumor tissues for molecular analyses. In the study information sheet given to the patients, nothing was written about hereditary breast cancer, and at study entry the patients did not expect to be contacted at a later point of time. Accordingly, patients were not biased to participate because of a wish to find out whether their breast cancer was of a hereditary type or not.

Five hundred and thirty-eight patients were included in the ABiM study. Prior to surgery, blood samples were taken and biobanked within 2 hours as whole blood, buffy coat, plasma, and serum similarly to as previously described [1]. At time of surgical removal, the tumor specimen was kept on ice and, after clinical pathological evaluation, unfixed fresh tissue (approximately 10–100 mg) was snap frozen and stored at –80°C. Tumor DNA/RNA was isolated using the AllPrep DNA/RNA Kit (Qiagen) and normal blood buffy coat DNA was isolated using the QIAamp DNA Blood Mini Kit (Qiagen), similar as previously described [1, 2].

No research tissue was taken unless it was certain not to influence the quality of diagnostic procedures. As a consequence, as well due to the quantity requirements of 10 μg tumor and 3 μg normal DNA, for the present study we analyzed 276 patients from the ABiM cohort. Three patients were excluded after quality control of the sequencing data (see below). The remaining 273 patients constitute our study population. Comparisons between the study population and the patients from the ABiM cohort that were not included in the present study population are presented in Supplementary Table S1. Compared to the patients not included, patients included in the study population were of similar age (62 vs. 63 years; $p = 0.29$), had larger tumors (≥ 20 mm: 47% vs. 38%; $p < 0.001$), of higher grade (Nottingham grade 3: 53% vs. 25%; $p < 0.001$), with higher Ki-67 (Ki-67 > 20: 46% vs. 23%; $p < 0.001$). Clinical information was retrieved from hospital records and INCA, the national cancer diagnosis quality registry. Ki-67 was measured as part of the SCAN-B project [1]. This study was conducted in accordance with the Declaration of Helsinki and has been approved by the Regional Ethical Review Board of Lund (diary numbers 2007/155, 2009/10, and 2009/658). Written information was given by trained health professionals and all patients provided written informed consent.

Statistical analyses

Differences in patient and tumor characteristics were tested using Fisher's exact test for categorical variables and Mann Whitney U test for continuous variables. All tests were two-tailed. All analyses were conducted using R version 3.1.

Targeted sequencing

Sequencing libraries were prepared from 278 tumor and 276 normal DNA samples from 276 patients (two patients had bilateral tumors) and enriched for targeted gene regions using a custom Agilent SureSelect target enrichment design with 120 bp tiled baits as previously described [3] (Supplementary Table S2). In brief, 2 µg DNA was sheared using Covaris ultrasonication and barcoded adapters ligated, amplified, and hybridized to the baits following the manufacturer's SureSelect Target Enrichment for the Illumina Paired-End Sequencing Library Protocol with some modifications (BGI Technologies): three or four barcoded samples were pooled for each hybridization reaction, and 13 cycles of post-hybridization PCR were performed. Libraries were sequenced on Illumina HiSeq 2000 instruments with paired-end 101 bp reads.

Sequence alignment and quality control

Illumina sequencing paired end reads were aligned to the human reference genome GRCh37 (downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz) using Novoalign v2.07.18 (Novocraft Technologies, Malaysia) with default settings. Duplicate read pairs were flagged with Picard MarkDuplicates v1.66 (Broad Institute, Cambridge, USA) and ignored in subsequent analyses. The median insert size was 211 bp (137 – 248 bp) and the *BRCA1* and *BRCA2* genes were sequenced to a median coverage of 603x (range 138 – 1541x) with 99.95% (range 97.7 – 100%) of bases having >30x coverage (Supplementary Table S3). Identity and matching between tumor and normal samples were confirmed by SNP analysis. After variant calling germline variants for each patient (see below), we excluded patient samples where the fraction of germline variants that was present in dbSNP was less than 75%. Three patients (tumor/normal) were excluded due to this, leaving 273 analyzed patients.

Variant calling

We used VarScan v2.3.5 [4] to call single nucleotide variants (SNVs) and indels in the *BRCA1* and *BRCA2* gene. To this end, we generated pileup files using samtools for each tumor and each normal sample Binary Alignment/Map (BAM) file using the following settings: *-B -Q 13 -q 1 -d 25000.* VarScan somatic was then run with the tumor and normal pileup file as input and the following settings: *--min-reads2 2 --min-avg-qual 15 --min-var-freq 0.05 --somatic_p_value 0.05 --strand-filter 1*. Next, the resulting variant calls were filtered with VarScan somaticFilter to remove clusters of false positive SNV calls near indels. The local reference sequence around each variant was retrieved using the getfasta command from the BEDTools suite [5]. Next, we ran bam-readcount (http://github.com/genome/bam-readcount) on each BAM file to generate metrics for further filtering. These metrics and the local sequence were used to remove potential false positive variant calls following the recommendations of the VarScan authors (Table 1 in [4]). Variants were classified as somatic if they were present in the tumor sample only, as germline if they were present in both the tumor and the normal sample, and were removed if they were present only in the normal sample.

Copy number analysis

CONTRA [6] was used to estimate DNA copy number for each sample. Tumor and normal samples were processed separately and compared against a virtual normal reference sample generated by pooling of all normal samples. Hetero- and homozygous deletions were called manually after visual inspection of the plotted copy numbers along all exons of *BRCA1* and *BRCA2* for each sample.

Annotation of variant calls

We used ANNOVAR [7] version 2014Jul22 to annotate the genomic positions of all variants with the ANNOVAR databases *refGene* (RefSeq transcripts), *snp138NonFlagged* (dbSNP build 138 without

SNPs having unknown or < 1% minor allele frequency, mapping more than only once to reference assembly, or flagged as clinically associated), and *ClinVar* [8]. To map genomic mutations to protein mutations we used RefSeq transcript NM_007294 for *BRCA1*, and NM_000059 for *BRCA2*. For each variant we also checked whether it had been submitted to the Breast Cancer Information Core (BIC, [9]).

Assessment of the deleteriousness of *BRCA1* and *BRCA2* variants

Variants located in introns (excluding splice sites) and synonymous SNVs were excluded. All variants (SNVs or indels) that resulted in a frameshift or a loss or gains of a stop codon were considered deleterious since they result in a nonsense or truncated protein sequence. Variants in the last exon of *BRCA2* were not considered deleterious. We considered SNVs that resulted in the change of one amino acid as deleterious if they were annotated as class 5 (pathogenic) in BIC or pathogenic in ClinVar or if Align-GVGD (http://agvgd.iarc.fr) predicted a class of C65 (deleterious).

Gene expression profiling

For all 276 breast tumors, RNA-sequencing libraries were generated from 1 μg tumor RNA and paired-end sequenced with 50 bp reads on an Illumina HiSeq 2000 using methods described elsewhere [1]. RNA-seq data was processed and gene expression counts were estimated using a Bowtie2/TopHat2/Cufflinks2 pipeline as described in [1]. Gene expression was summed by collapsing on unique gene symbols.

Intrinsic breast cancer subtype

Tumors were subtyped according to St. Gallen criteria as well as by PAM50 gene expression subtyping. For St. Gallen subtyping, we assigned each tumor to one of the following five subtypes [10]: Basal, if tumor is triple negative (ER negative, PR negative, and HER2 negative); Luminal A, if tumor is ER or PR positive, HER2 negative, and Ki-67 ≤ 20; Luminal B HER2–, if tumor is ER or PR positive, HER2 negative, and Ki-67 > 20; Luminal B HER2+, if tumor is ER or PR positive, HER2 positive, and Ki-67 > 20; Non-luminal HER2+, if tumor is ER and PR negative, and HER2 positive. ER and PR status was assessed using immunohistochemistry (IHC) with a positive tumor defined as having ≥ 1% cells stained positive. HER2 positive tumors were either IHC 3+ or fluorescence in situ hybridization (FISH) positive, and HER2 negative tumors were either IHC 0 or 1+ or FISH negative. PAM50 subtyping was performed using an implementation of the Parker method [11]. In short, to avoid context dependency when assigning PAM50 subtype by nearest-centroid, a fixed reference was selected to match the original cohort used by Parker *et al.* with respect to available clinical characteristics. Before subtyping tumors in this study, gene expression of the PAM50 genes for each tumor was centered to the reference set separately using custom R scripts.

Survival analysis

For overall survival (OS), vital status was checked in the Swedish Census Register. For recurrence-free survival (RFS), recurrence information was obtained from the clinical cancer database INCA. Events were death of any cause for OS, and local or distant recurrence for RFS. Times were measured from date of diagnosis to date of either an event (see above), death, or last follow-up (whatever occurred first). Times were censored if no event occurred, and uncensored otherwise. OS and RFS were estimated using the Kaplan-Meier method and groups were compared using the log-rank test (two-tailed).

References

1. Saal LH, Vallon-Christersson J, Häkkinen J et al. The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med 2015; 7: 20.
2. Olsson E, Winter C, George A et al. Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic disease. EMBO Mol Med 2015; 7: 1034-1047.
3. Harbst K, Lauss M, Cirenajwis H et al. Molecular and genetic diversity in the metastatic process of melanoma. J Pathol 2014; 233: 39-50.
4. Koboldt DC, Zhang Q, Larson DE et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012; 22: 568-576.
5. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010; 26: 841-842.
6. Li J, Lupat R, Amarasinghe KC et al. CONTRA: copy number analysis for targeted resequencing. Bioinformatics 2012; 28: 1307-1313.
7. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010; 38: e164.
8. Landrum MJ, Lee JM, Riley GR et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014; 42: D980-985.
9. Szabo C, Masiello A, Ryan JF, Brody LC. The breast cancer information core: database design, structure, and scope. Hum Mutat 2000; 16: 123-131.
10. Goldhirsch A, Wood WC, Coates AS et al. Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol 2011; 22: 1736-1747.
11. Parker JS, Mullins M, Cheang MC et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009; 27: 1160-1167.